



**Computational genomic analyses of long-lived mammals to study  
variation in cancer resistance, longevity and life history**

**Thesis submitted in accordance with the requirements of the  
University of Liverpool for the degree of Doctor in Philosophy**

**by**

**Michael Keane**

**February 2018**

## Abstract

Little is known about the genetic and molecular mechanisms responsible for the great differences in mammalian longevity and life history. One potential source of novel insights is based on comparative analyses of the genomes of species which exhibit extreme longevity and an extended life history. As such, this work describes the results obtained from the analysis of the bowhead whale, naked mole rat (NMR) and human genomes, each of which are exceptionally long-lived compared to closely-related species.

The bowhead whale genome was analysed with a focus on identifying genes with evidence of positive selection and proteins with unique amino acid residues when compared with other mammals. A number of genes that have previously been associated with cancer and ageing were found to exhibit evidence of positive selection on the bowhead lineage. In addition, bowhead-specific alterations in proteins linked to sensory perception of sound and size and development were also identified which are of potential relevance due to the phenotypic divergence of the bowhead whale associated with these traits.

The analysis of the NMR assembly attempted to identify genes with a signal of positive selection by comparing synonymous and non-synonymous substitution rates. While positive selection on NMR genes has previously been analysed, we found additional signals of selection in several which have not previously been reported, including in regions of p53 and the hyaluronan receptors *CD44* and *HMMR*.

Finally, while the previous analyses focused on coding sequences, it is also likely that much of the genetic basis for the variation in longevity is to be found in non-coding regions of the genome. In order to assess this hypothesis, human data from both genome wide association studies (GWAS) and annotated 3'UTR sequences was analysed in order to identify genes with signals of molecular adaptation which correlate with trait divergence. The GWAS meta-analysis identified genes from a specific pathway which has previously been shown to regulate the timing of growth and development. The genes identified in the 3'UTR analysis were slightly below the level of statistical significance indicating that greater statistical power, most likely in the form of including sequences from additional species, is necessary.

Overall, the results obtained offer novel insights regarding the molecular adaptations by which longevity and life history evolve and identify numerous genes which could be prioritised for future studies including potential functional characterisation. Furthermore, all the data and results

generated have been made available on customised online portals in order to allow easy access to the scientific community and facilitate further research into these long-lived species.

## **Acknowledgements**

I would like to acknowledge the supervision during the writing of this thesis of Prof. Andy Jones, who provided invaluable input and support throughout the process. I would also like to acknowledge Dr. João Pedro de Magalhães for providing the opportunity to participate in several genome sequencing and analysis projects, and thank the members of his group, particularly Daniel for his hospitality and Sipko for the comedy. This work was supported by a studentship from the University of Liverpool's Faculty of Health and Life Sciences.

## Table of Contents

Chapter 1	Introduction and aims	6
Chapter 2	Annotation and analysis of the bowhead whale genome	28
Chapter 3	The Naked Mole Rat Genome Resource: facilitating analyses of cancer and longevity-related adaptations	61
Chapter 4	MYCN/LIN28B/Let-7/HMGA2 pathway implicated by meta-analysis of GWAS in suppression of post-natal proliferation thereby potentially contributing to aging	65
Chapter 5	Systematic detection of positive selection on human 3'UTRs	69
Chapter 6	Discussion and conclusions	75
	References	79
	Appendices	89

# **1. Introduction and aims**

## **1.1 Mammalian life history**

There is enormous variation in mammalian life history (the sequence of events related to survival and reproduction that occur from inception to death) and longevity (herein defined as maximum lifespan), even in relatively closely related species (de Magalhães et al., 2007). By way of illustration, mice (*Mus musculus*) reach maturity at 42 days and can live for 4 years, compared with 228 days and 31 years respectively for the naked mole rat (*Heterocephalus glaber*), a rodent of similar size, and great variation in lifespan is evident even among a subset of species for which genome sequences are available (Table 1.1).

Table 1.1: maximum lifespans of selected mammalian species with sequenced genomes  
(<http://genomics.senescence.info/species/>).

Species	Name	Maximum lifespan (years)
Human	<i>Homo sapiens</i>	122.5
Chimpanzee	<i>Pan troglodytes</i>	59
Gorilla	<i>Gorilla gorilla</i>	55.4
Orangutan	<i>Pongo pygmaeus</i>	59
Rhesus monkey	<i>Macaca mulatta</i>	40
Marmoset	<i>Callithrix jacchus</i>	16.5
Tarsier	<i>Tarsius syrichta</i>	16
Bushbaby	<i>Otolemur garnettii</i>	18.3
Mouse lemur	<i>Microcebus murinus</i>	18.2
Mouse	<i>Mus musculus</i>	4
Rat	<i>Rattus norvegicus</i>	5
Guinea pig	<i>Cavia porcellus</i>	12
Squirrel	<i>Spermophilus tridecemlineatus</i>	7.9
Kangaroo rat	<i>Dipodomys ordii</i>	9.9
Naked mole rat	<i>Heterocephalus glaber</i>	31
Cow	<i>Bos Taurus</i>	20
Dolphin	<i>Tursiops truncatus</i>	51.6
Orca	<i>Orcinus orca</i>	90
Minke whale	<i>Balaenoptera acutorostrata</i>	50
Bowhead whale	<i>Balaena mysticetus</i>	211

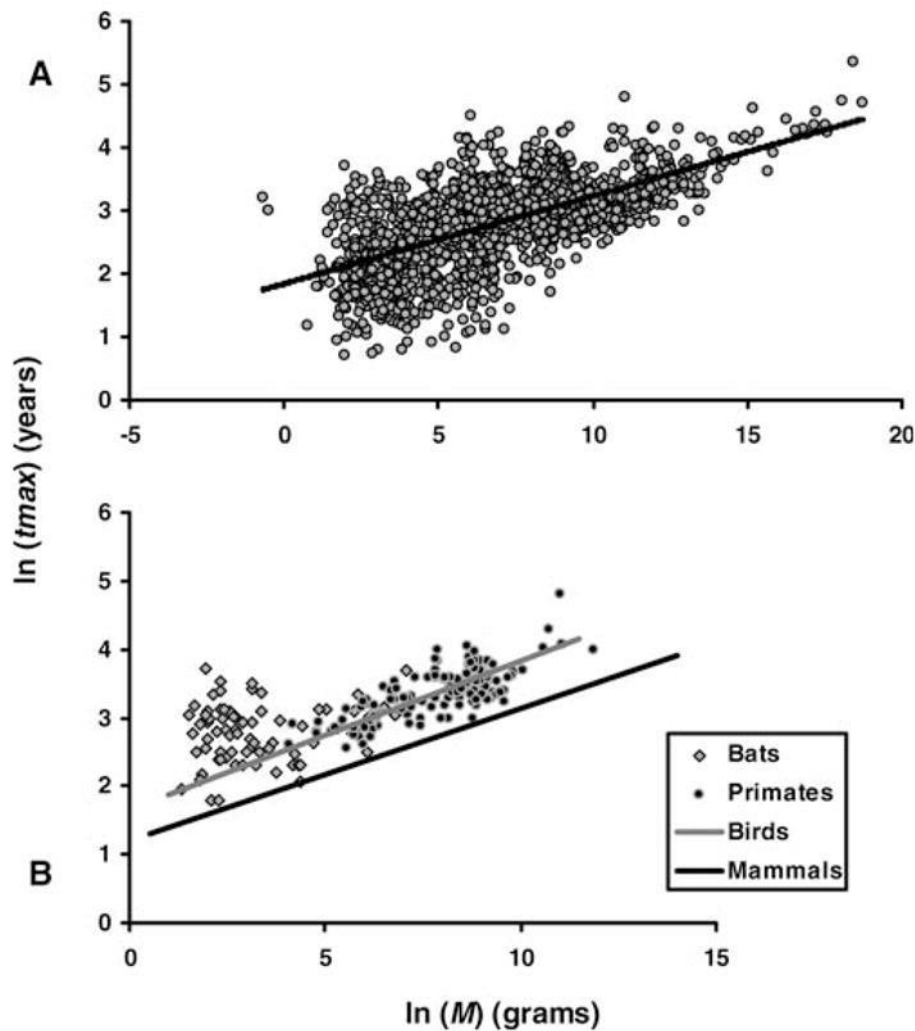
It is important to take account of potential ascertainment biases in relation to maximum lifespan as presented above. For example, there is significant variation in sample sizes across species which could contribute to inaccurate estimation. In particular, the sample size for human vastly exceeds that of any other species and as such, human maximum longevity relative to other species is most likely significantly overstated. An additional issue is that for some species, it is difficult or impossible to measure longevity directly. This particularly applies to species such as the bowhead whale, for which it is necessary to estimate longevity indirectly using a method known as aspartic acid racemization (Garde et al., 2007), which cannot be expected to have the same accuracy as direct observation.



## 1.2 Maximum lifespan

Maximum lifespan strongly correlates with a number of factors, most notably body mass (Figure 1.1), with larger species living longer than smaller species on average (Promislow, 1993).

Figure 1.1: plot of the ln-transformed relationship between body mass ( $M$ ) and maximum longevity ( $t_{\max}$ ) across vertebrates (de Magalhães et al., 2007). **A**, Grey circles: all mammal, bird, reptile and amphibian species ( $n = 1456$ ) in AnAge (de Magalhães et al., 2005). **B**, Grey line: avian regression curve; black line: mammalian regression curve minus bats and cetaceans. Closed circles: primates ( $n = 137$ ); grey squares: bats ( $n = 73$ ), the two longest-lived mammalian orders for their body size.



However there are several exceptions to this tendency, in particular birds and bats (Austad and Fischer, 1991). Ecological constraints offer a plausible explanation reconciling these observations. Flight and increased body size contribute to a reduced risk of predation, leading to a lower mortality rate, higher maximum lifespan and reduced rate of aging, influencing the evolution of life history and longevity (Stearns, 1992).

### **1.3 Adaptation**

When considering the evolution of any phenotypic trait, a clear understanding of the concept of adaptation is essential. Adaptation is defined as the evolutionary process by which organisms become better suited to life in their habitat or habitats (Dobzhansky et al., 1968). The prevailing view had originally been that natural selection (Darwin, 1859), or the differential inheritance of alleles associated with beneficial phenotypic variation, was the dominant mechanism underlying adaptation at the molecular level. However this was challenged by the neutral theory of molecular evolution (Kimura, 1968), which posits that the majority of the molecular variation within and between species is caused not by natural selection, but by random genetic drift of neutral alleles.

### **1.4 Genetic drift**

This sparked an extended debate regarding the relative influence of natural selection and genetic drift (Wright, 1929), which is the change in the frequency of a gene variant in a population due to random sampling, on molecular evolution. A number of mathematical models have been created to describe changes in allele frequency in an idealised population, which support the Hardy-Weinberg principle (Hardy, 1908; Weinberg, 1908). This states that within sufficiently large populations, the allele frequencies remain constant from one generation to the next unless the equilibrium is disturbed by external influences such as migration, mutation or selection. As such, the effective population size would be expected to have a significant impact on the influence of genetic drift relative to natural selection. As a simple illustration, whereas in an infinite hypothetical population fixation of a specific allele would be impossible, as effective population size decreases, sampling error can result in significant alteration of allele frequencies.

## 1.5 Genetic basis of longevity

Although there is clearly a genetic effect in terms of variation in lifespan, little is currently known about the genomic elements responsible (Finch, 1990). It has been proposed that Apolipoprotein E (APOE), involved in age-related diseases and associated with human longevity, is a meat-adaptive protein that contributed to the evolution of human lifespan (Finch and Stanford, 2004). Specifically, carriers of the  $\epsilon 4$  allele of the APOE gene (responsible for lipid transport) have higher levels of total cholesterol and accumulation of atherosclerotic plaques in arteries, leading to increased risks of cardiovascular disease and stroke, as well as dementia and AD (Fullerton et al., 2000). The APOE  $\epsilon 2$  and  $\epsilon 3$  alleles confer reduced risks of these diseases of aging relative to the  $\epsilon 4$  allele and are relatively recent additions to the human genome, with the  $\epsilon 3$  and  $\epsilon 2$  allele clade having evolved by 200,000 years ago (Fullerton et al., 2000). Today, prevalence of these alleles varies around the world but in most populations,  $\epsilon 3$  is found in the highest frequency (mean = 78.3%; range: 8.5–98.0%), followed by  $\epsilon 4$  (mean = 14.5%; range: 0–49%) and  $\epsilon 2$  (mean = 6.4%; range: 0–37.5%) (Eisenberg et al., 2010). Therefore it was proposed that humans' exceptionally long lifespans are a product, in part, of the evolution of the  $\epsilon 3$  allele, especially because diets shifted to include more meat and increased dietary fat and cholesterol later during our evolutionary history (Finch and Stanford, 2004).

A small number of studies have attempted to identify coding genes associated with the evolution of mammalian longevity. The first searched for codons and genes showing a stronger level of amino acid conservation in long-lived than in short-lived lineages (Jobson et al., 2010), however this approach appears counter-intuitive if longevity increased as mammals evolved from the ancestral lineage. Another analysis scanned for longevity-selected positions in the mammalian proteome, and reported several proteins that interact in inflammation and other aging-related processes, as well as in organismal development (Semeiks and Grishin, 2012). A third study identified proteins which had evolved faster on long-lived lineages (Li and de Magalhães, 2013), however it is based on numerous assumptions about the branches of the mammalian phylogeny on which longevity increased, which may require justification. In spite of this work, to date there is no example of a functionally validated gene associated with variation in mammalian longevity.

## 1.6 Comparative genomics

One potential means of attempting to isolate the molecular basis of such complex traits is to compare and analyse the genomes of closely-related species and identify genetic differences which could plausibly be responsible. Such comparative genomics approaches exploit the trait divergence between species to find correlated genomic differences. The opportunities to apply comparative methods continue to grow as the list of species with sequenced genomes has expanded greatly in recent years, primarily due to the continuously declining cost of sequencing, and now includes bat (Zhang et al., 2013), pigeon (Shapiro et al., 2013), turtle (Shaffer et al., 2013), cobra (Vonk et al., 2013), camel (Jirimutu et al., 2012), tiger (Cho et al., 2013), orca (Foote et al., 2015), minke whale (Yim et al., 2014) and bowhead whale (Keane et al., 2015).

A number of intriguing results have already been reported following the comparative analyses of these genomes. For example, it was reported in the analysis of two bat genomes that a high proportion of genes in the DNA damage checkpoint–DNA repair pathway, including ATM, TP53, RAD50 and KU80, are under selection (Zhang et al., 2013). This is notable because these genes have been directly associated with ageing in model systems and therefore, it points towards a potential role for averting DNA damage in longevity assurance mechanisms; a notion dating back several decades that remains contentious. The analysis of the pigeon genome identified the gene EphB2 as a strong candidate for the derived head crest phenotype shared by numerous breeds, an important trait in mate selection in many avian species (Shapiro et al., 2013). Rapidly evolving genes in the camel lineage are significantly enriched in metabolic pathways, which may underlie the insulin resistance typically observed in these animals (Jirimutu et al., 2012). Finally, the comparative analysis of the genomes of the killer whale and other marine mammals found that convergent amino acid substitutions were widespread throughout the genome and that a subset of these substitutions were in genes evolving under positive selection and putatively associated with a marine phenotype (Foote et al., 2015).

These analyses did not however extend beyond the examination of protein-coding sequences, which it should be borne in mind constitute only a minority of the constrained sequence in the genome (Rands et al., 2014). While regulatory sequences are more difficult both to identify and analyse, significant progress has also been made in this area. Early studies identified hundreds of so-called ultra-conserved elements—elements several hundred bases long and almost identical across mammals (Bejerano et al., 2004)—and functional studies demonstrated that some of these had a role as enhancers (Woolfe et al., 2005). Comparison of human, mouse, rat and dog identified several hundreds of thousands of conserved non-coding elements that cluster near developmental genes

(Lindblad-Toh et al., 2005), suggesting the importance of gene regulation for determining body plan and neurological development. More recently, an analysis of 29 mammalian genomes identified 3.6 million constrained elements encompassing 4.2% of the human genome (Lindblad-Toh et al., 2011).

A lineage-specific alteration in an otherwise well-conserved element is an important indication of innovation and as such, the identification of a set of elements that are highly-conserved across vertebrates but exhibit accelerated evolution in humans (Pollard et al., 2006) was of particular significance. In addition, conserved non-coding elements in mammals may also be deleted in humans and close relatives (McLean et al., 2011). Further studies of natural selection, notably in sticklebacks (Jones et al., 2012), clearly demonstrate the contribution of regulatory innovation to phenotypic evolution.

## 1.7 Detection of positive selection

In the context of the comparative analysis of selective forces, a very commonly-used approach involves comparing the rate of non-synonymous (dN) to synonymous (dS) substitutions per respective site in protein-coding genes. Non-synonymous mutations result in changes in the amino acid sequence of the protein while synonymous mutations do not. This method has been frequently applied to detect selective pressure acting on genes in genome-wide surveys (Chimpanzee Sequencing and Analysis Consortium, 2005).

In the classical test,  $dN/dS < 1$  indicates purifying selection,  $dN/dS = 1$  is indicative of neutral selection and  $dN/dS > 1$  is suggestive, but not necessarily proof, of positive selection. There are some obvious issues with this simple approach. Firstly, it is possible for  $dN/dS$  to exceed 1 purely by chance as a result of random fluctuations in dN and/or dS. Therefore it is essential to also estimate the statistical significance of the  $dN/dS$  values calculated and only predict positive selection when  $dN/dS > 1$  with a significant p-value. In addition, positive selection may only act on a small number of sites over a short period of evolutionary time, resulting in any signal being swamped by the ubiquitous purifying selection on protein-coding genes. A branch-site method was therefore developed in order to detect positive selection on individual codons along specific lineages (Yang and Nielsen, 2002), which addresses these issues and allows an assessment of the statistical significance.

Using this method, branches of the phylogenetic tree are divided *a priori* into foreground and background lineages, and a likelihood ratio test is constructed by comparing a model that allows positive selection on the foreground lineages with a model that does not allow such positive selection. However this test was found to generate an excessive level of false positives if some sites evolve under negative selection on the background lineages, but experience a relaxation of constraints on the foreground lineages (Zhang 2004). To address this issue, a slightly modified version was introduced which used a Bayesian approach and assigned a prior to the model parameters, which in comparison with the original method was found to alleviate the problem of false positives (Yang et al., 2005).

When considering the results of any genome-wide survey of positive selection, there are a number of issues that should be borne in mind in relation to confounding factors which could affect the accuracy of predictions. In particular, sequencing, annotation and alignment errors (Schneider et al., 2009) as well as recombination (Anisimova et al., 2003) have significant potential to inflate estimates of selective pressure. Using the human genome as a point of reference, published estimates of the proportion of positively selected genes (PSGs) range between 0.02% (Gibbs et al., 2007) and 8.7%

(Clark et al., 2003). It is plausible that a proportion of the enormous variation between these estimates is due to errors in the methodologies used. In addition, there have been some surprising and potentially dubious claims regarding variation in the extent of positive selection between close relatives, including that the proportion of PSGs in human is only half the number in chimpanzee (Bakewell et al., 2007).

To assess the effect of errors, almost 3,000 orthologous protein-coding genes were used to infer the fraction of PSGs in seven terminal mammalian branches (Schneider et al., 2009). It was observed that the quality of the sequence, the degree of mis-annotation and ambiguities in the multiple sequence alignment had a significant impact on the proportion of genes exhibiting signals of positive selection. In particular, the inferred fraction of PSGs in sequences that were deficient in each of coverage, annotation and alignment was 7.2 times higher than that in genes with high trace sequencing coverage, “known” annotation status and perfect alignment scores. Of these three issues, sequence quality had the greatest impact, with the proportion of PSGs in low coverage sequences found to be 3.3 times that in high coverage sequences.



## 1.8 Genome sequencing and assembly

Sequence coverage refers to the average number of reads per locus and is of great significance as an indicator of genome sequence quality. To illustrate, in a genome sequencing project in which shotgun sequencing strategy is used, genomic DNA is firstly sheared into small random fragments. Depending on the sequencing platform used, these fragments are then sequenced independently to a given length. Commonly used platforms at present include short-read technologies such as Illumina HiSeq (typically 150 bp) and SOLiD (typically 50 bp), which have become popular alternatives to traditional Sanger sequencing (~1 kb) and Roche 454 sequencing (up to 800 bp). Whole genome sequencing using these platforms requires a significant amount of high-quality, non-degraded DNA (Wong et al., 2012), which can be a significant issue for species with conservation concerns.

The resulting sequence reads are then assembled back together into longer continuous stretches of sequence known as contigs by powerful computer software. ALLPATHS-LG (Gnerre et al., 2011) and SOAPdenovo (Luo et al., 2012) are two commonly-used software applications employed for this purpose, which is known as *de novo* genome assembly. After the initial assembly, contigs are typically joined to form longer stretches of sequence known as scaffolds. To do so, paired-end libraries of long DNA fragments are prepared and their endpoints sequenced. The endpoint sequences of independent fragments which lie on two different contigs are joined into a scaffold.

There are a number of confounding issues that can arise during this process. Firstly, if there is not sufficient overlap between the sequence reads at each position on the genome, difficulties can arise in relation to the assembly of the reads into contigs at these locations. Higher sequence coverage or longer sequence reads are possible means by which this issue can be addressed. In addition, *de novo* assemblies generated using short sequence reads encounter significant issues with both repetitive and duplicate sequences. This is because the reads typically do not extend beyond the repeat and duplicate regions, with the result that the assembly software cannot distinguish between them (Alkan et al., 2011). To illustrate the magnitude of this issue, it was reported that *de novo* assemblies were 16.2% shorter than the reference genome, with 420.2 megabase pairs of common repeats, 99.1% of validated duplicate sequences and over 2,377 coding exons missing from the genome (Alkan et al., 2011).

## 1.9 Gene prediction and annotation

Genome annotation and gene prediction is the process of identifying genes and their intron–exon structures in a genome assembly. Therefore the first step in this process is clearly identifying whether a genome assembly is suitable for annotation. A commonly-used statistic for this purpose is N50. Given a set of scaffolds, each with its own length, the N50 is the length for which the collection of all scaffolds of that length or longer contains at least half of the sum of the lengths of all scaffolds, and for which the collection of all scaffolds of that length or shorter also contains at least half of the sum of the lengths of all scaffolds. Although there is no agreed standard, an assembly with an N50 equal to or greater than the median gene length is typically regarded as a good candidate for annotation (Cantarel et al., 2008; Ye et al., 2011). Because N50 is calculated in the context of the assembly size rather than the genome size, comparisons of N50 values derived from assemblies of significantly different lengths are usually not informative, even for the same genome. To address this issue, an alternative statistic known as the NG50 can be used. This is identical to the N50 except that it is 50% of the known or estimated genome size that must be of the NG50 length or longer, which allows for meaningful comparisons between different assemblies.

CEGMA (Parra et al., 2007) provides an additional means of estimating the completeness and contiguity of an assembly. This tool screens an assembly against a collection of essentially universal eukaryotic single-copy genes and also determines the percentage of each gene lying on a single scaffold. If an assembly is incomplete or if its N50 scaffold length is too short, it is advisable to complete additional shotgun sequencing, as tools are available for the incremental improvement of draft assemblies (Tsai et al., 2010; Assefa et al., 2009; Husemann and Stoye, 2010).

The next step is to actually annotate the assembly, which is done using software known as genome annotation pipelines. Although there are details specific to each pipeline, they share a core set of features. Generally, genome-wide annotation of gene structures is divided into two distinct phases. In the first, the ‘computation’ phase, evidence such as expressed sequence tags (ESTs), proteins and RNA-seq data are aligned to the genome and *ab initio* and/or evidence-driven gene predictions are generated. In the second or ‘annotation’ phase, these data are synthesized into gene annotations. Because this process is intrinsically complicated and involves many different tools, the programs that assemble the data and use it to create genome annotations are generally referred to as annotation pipelines. Current pipelines are focused on the annotation of protein-coding genes, although Ensembl (Flicek et al., 2013) also has some capabilities for annotating non-coding RNAs (ncRNAs).

The computation phase typically involves repeat identification, evidence alignment and gene prediction. Eukaryotic genomes can contain large amounts of repetitive sequence with 47% of the

human genome for example estimated to consist of repeats (Lander et al., 2001). Furthermore, repeats are often poorly conserved and therefore creation of a repeat library for the genome of interest is advisable, which can be used with a tool such as RepeatMasker (Smit and Hubley, 2011) to identify blocks of sequence in a target genome that are homologous to known repeats. Following repeat masking, most pipelines align evidence such as proteins, ESTs and RNA-seq data to the genome assembly. Because proteins retain substantial sequence similarity over far greater spans of evolutionary time than nucleotide sequences, proteins from other species are generally included also. UniProtKB/SwissProt (Bairoch et al., 2004) is an example of an excellent resource for protein sequences to be used in this process. The alignments generated are then generally 'polished' using tools such as Exonerate (Slater and Birney, 2005) in order to precisely define splice sites and exon boundaries, which allows the genes in the assembly to be predicted.

The annotation phase takes the results of these computations and attempts to synthesise them into gene annotations. Traditionally this was done by manual human review of the evidence for each gene to decide on the intron-exon boundaries, however this is so labour-intensive that genome projects generally rely on automated annotations. One approach is to feed the alignment evidence to gene predictors at run time (evidence-driven prediction) in order to improve the accuracy of the prediction process, and then identify the most representative prediction. This is the process used by MAKER (Holt and Yandell, 2011), which combines *ab initio* and homology-based methods to derive gene models also incorporating *de novo* prediction tools including BLASTX, Exonerate, SNAP, Genemark and Augustus.

Following completion of the process of annotation, it is critical to be aware of the obvious potential for poor quality annotation and gene prediction to seriously confound any downstream comparative genomics analysis. In this regard, it is interesting to note that even as the cost of genome sequencing has continued to fall, gene prediction has perhaps become more challenging. Indeed, it is rare for the accuracies of even the best genome annotation pipelines to exceed 80% at the exon level (Reese and Guigo, 2006). There are a number of confounding issues that contribute to explain this. Firstly, many recent sequencing projects focus on relatively exotic genomes without the availability of a high-quality annotation from a closely-related species, resulting in significant potential for poor-quality gene predictions. In addition, there is the increasing quantity, complexity and variety of data that must be merged into a single gene model, including protein, expressed sequence tag (EST) and more recently RNA-seq data both from the species of interest and related species. Another significant issue is that mis-annotations in any of these datasets, which can be quite common, have the obvious potential to confound gene predictions. Finally, the shorter read lengths generated by

current sequencing techniques can result in significantly decreased contiguity relative to traditional shotgun methods, increasing the complexity of generating gene models.

## 1.10 Biased gene conversion

In addition to sequencing and annotation errors, biased gene conversion (BGC) is another factor that has significant potential to contribute to inflated estimates of positive selection. BGC is a neutral recombination-associated process caused by the GC-biased repair of A:C and G:T mismatches in heteroduplexed recombination intermediates, which from a population genetics perspective appears equivalent to directional selection (Nagylaki, 1983). BGC results in highly recombining chromosomes and regions becoming rapidly GC-enriched (Galtier et al., 2001) to the extent that purifying selection can be overpowered, a spectacular example of which is provided by the *Mid1* gene in the mouse (Perry and Ashworth, 1999). This gene is X-linked in human, rat and short-tailed mouse *Mus spretus* but is translocated in *Mus musculus*, overlapping the boundary between the X-specific region and the pseudoautosomal region (PAR) such that the 5' end (exons 1-3) is in the X-specific region while the 3' end (exons 4-10) is within the PAR (Perry and Ashworth, 1999). The PAR is a short, highly recombining region of homology between the X and Y chromosomes (Soriano et al., 1987).

Therefore as a consequence of this translocation, the 3' *Mid1* sequence experienced a huge increase in recombination rate in *Mus musculus*. This resulted in a dramatic increase in GC content and substitution rate at third codon positions and introns at the 3' but not the 5' end of the gene (Perry and Ashworth, 1999). The *Mid1* protein is highly conserved with the human and *Mus spretus* sequences differing by only six amino acids. Remarkably however, there are 28 differences between the *Mus musculus* and *Mus spretus* proteins, all of which are due to AT → GC substitutions located in the 3' end of the gene. Clearly this cannot be explained by adaptation, as third codon positions and introns should then have been relatively unaffected. It is theoretically possible that it could be the result of selection favouring increased GC content to modulate gene stability (Jabbari and Bernardi, 2004) or expression level (Kudla et al., 2006), however in this case both the 5' and 3' ends of the gene should have been affected equally. This suggests that the accelerated evolution of *Mid1* in *Mus musculus* is the result of BGC and clearly illustrates how it can result in inflated estimates of positive selection.

Taken together, these findings indicate that computational analyses of selection should be stringently assessed for confounding factors before they are regarded with confidence.

### **1.11 Evolution of protein sequences**

In relation to the evolution of protein sequences, it is possible to identify highly conserved amino acid residues that are uniquely altered in a species of interest (Kim et al., 2011), which has been shown to be indicative of alterations in protein function (Tian et al., 2013). Identification of gene duplications can also be of relevance in terms of phenotypic variation (Holland et al., 1994). These techniques were employed in this work for the analysis of the genomes of the bowhead whale (Chapter 2) and naked mole rat (Chapter 3).

## 1.12 Genome-wide association study design

It is generally accepted that variation between species emerges from variation within a species. While the methods discussed thus far have been used to analyse molecular evolution between species, genome-wide association studies (GWAS), which use genetic variation across the genome to associate genetic variation with phenotypic divergence (Hunter et al., 2008), have been widely applied using within-species data, human in particular.

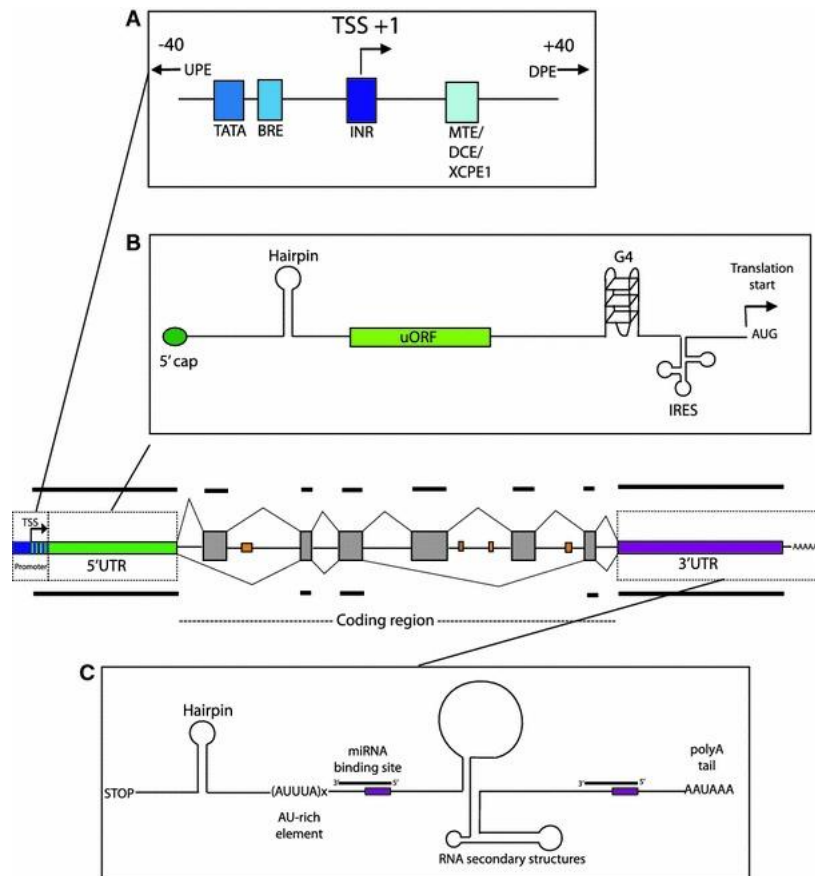
In terms of study design, the most commonly used approach, known as case-control, is based on identifying genomic loci containing single nucleotide polymorphisms (SNPs) which are significantly enriched in a case cohort exhibiting a specific phenotype, for example a disease, relative to a control cohort. A fundamental concept in the context of GWAS is linkage disequilibrium, which is the non-random association of alleles at linked loci. This allows the genotyping of only a subset of the known SNPs, which can be used to impute SNPs that were not genotyped (Marchini et al., 2007). An odds ratio and p-value are then computed for each SNP. The odds ratio quantifies the extent to which a specific allele is present or absent in the case relative to the control cohort, and a chi-squared test is typically used to calculate the p-value for the significance of the odds ratio. A Manhattan plot can be generated using these results, illustrating the negative logarithm of the p-value as a function of genomic location. Because of the very large number of SNPs tested in a GWAS, the p-value must be corrected for multiple testing issues. As a result, the threshold for significance is often in the region of  $p < 5e-8$ . The GWAS approach has already been used to study genetic variation associated with human longevity, but unfortunately this high-profile study was subsequently retracted due to technical errors and an inadequate quality control procedure (Sebastiani et al., 2011), and the only high-confidence finding in the republished analysis (Sebastiani et al., 2012) was *APOE*, which is already known. However the results of a large number of GWAS on longevity and other traits have now been catalogued, which creates the opportunity for meta-analysis of traits related to life history and potentially longevity (Chapter 4).

### **1.13 The regulatory landscape of genes**

Even though typically only coding sequences are annotated for a newly sequenced genome, it should be borne in mind that significantly more detail is required before a gene annotation can be considered complete. In particular, the regulatory context of the gene should also be considered. In the first instance this is in reference to proximal regulatory elements, such as untranslated regions (UTRs) and promoters. The promoter is located upstream of the transcription start site and is primarily responsible for initiation of gene transcription (Smale and Kadonaga, 2003). Promoters contain binding sites for RNA polymerase and transcription factors that play a regulatory role during transcription. The 5' UTR immediately precedes the initiation codon and following transcription into mRNA, plays an important role in the regulation of translation by differing mechanisms in eukaryotes and prokaryotes (Bannerjee, 1980). The 3' UTR lies directly downstream of the termination codon and typically contains regulatory elements that modulate gene expression levels and in particular, binding sites for microRNAs (miRNAs) and regulatory proteins (Hesketh, 2004). Decreased expression of mRNAs can be effected by miRNAs binding to complementary sites in the 3' UTR, causing degradation or reduced translation of the transcript (Barrett et al., 2012). 3' UTRs can also contain AU-rich elements (AREs), which are typically 50 to 150 bases in length and can alter expression levels when bound by ARE binding proteins. There are also longer-range regulatory elements that can allow activation or inhibition of transcription, including enhancers, insulators and silencers (Barrett et al., 2012). Enhancers contain binding sites which when bound by the complementary transcription factors serve to activate transcription of a gene by interaction with the promoter, whereas insulators block this interaction and have an inhibitory effect on transcription. Finally transcription can also be inhibited by repressors which by binding to silencers prevent RNA polymerase from binding to the promoter.



Figure 1.2: Regulatory elements in noncoding gene regions (Barrett et al., 2012). The centre image shows a typical gene, with exons indicated in grey. The orange rectangles indicate intronic enhancer elements. a. Promoter region regulatory elements. Upstream and downstream promoter elements situated outside of the core promoter region are indicated by the arrows. b. Regulatory elements in the 5'UTR. c. Regulatory elements in the 3'UTR.



### **1.14 Selection on proximal non-coding sequence**

Although there have been many analyses of positive selection on coding genes, these have thus far failed to identify the genomic basis of the most complex and interesting trait adaptations, including encephalisation, skeletal morphology and longevity in the case of human evolution (Varki et al., 2008). By contrast, little is known of the extent of selection on functional non-coding sequence (Carroll, 2005). While a small number of studies have identified conserved non-coding regions displaying accelerated rates of evolution on the human lineage (Pollard et al., 2006; Lindblad-Toh et al., 2011), it is unclear which genes are subject to regulation by these elements. This is not surprising given that the regulatory targets of the vast majority of conserved non-coding sequences, in particular developmental enhancers, are unknown (Lettice et al., 2003). This limitation could potentially be avoided by focusing on proximal non-coding DNA, i.e. UTRs and promoters, as the regulated genes are evident. To assess the efficacy of this hypothesis, an exploratory analysis of selection on human 3' UTRs was completed (Chapter 5).

## 1.15 Aims

The genomic adaptations responsible for variation in life history, longevity and resistance to age-related diseases in mammals are largely unknown. Some progress has been made in model organisms, most notably *C. elegans*, in which hundreds of genes associated with the modulation of longevity are now known (<http://genomics.senescence.info>), and a network of transcription factors and microRNAs regulating the progression through the stages of its life history, including *LIN-28* and *let-7*, has been characterised (Ambros, 2011). However the extent to which findings in relatively short-lived model organisms are applicable to long-lived mammals, and humans in particular, is debateable. Indeed, there has to date been little research focus on molecular evolution in particularly long-lived mammals. As such, the primary aim is to apply comparative genomic methods, as outlined above, to genomic data from species which exhibit exceptional longevity in order to derive insights into the molecular adaptations responsible for the evolution of an extended life history and longevity in mammalia.

In addition, because cancer has been investigated to a significantly greater degree than longevity and a large number of genes have been associated, a secondary objective is to assess the degree of molecular adaptation in cancer-associated genes. Although larger species paradoxically tend to experience lower cancer incidence compared to smaller species despite having a greater cell count, the molecular basis of this observation remains unknown (Caulin and Maley, 2011). It should be borne in mind however that an animal's cell count is a function not just of the rate of cell division, but also cell death. It is thus the combination of these factors that ultimately determines body mass. In this context, it is clear that species differences in the rates and distribution of programmed cell death may also have a significant effect, however this has not yet been systematically assessed (Kuida et al., 1998).

The following chapters therefore describe the application of a number of comparative genomics methods to genomic data from notably long-lived mammalian species, specifically bowhead whale (Chapter 2), naked mole rat (Chapter 3) and human (Chapters 4 and 5).

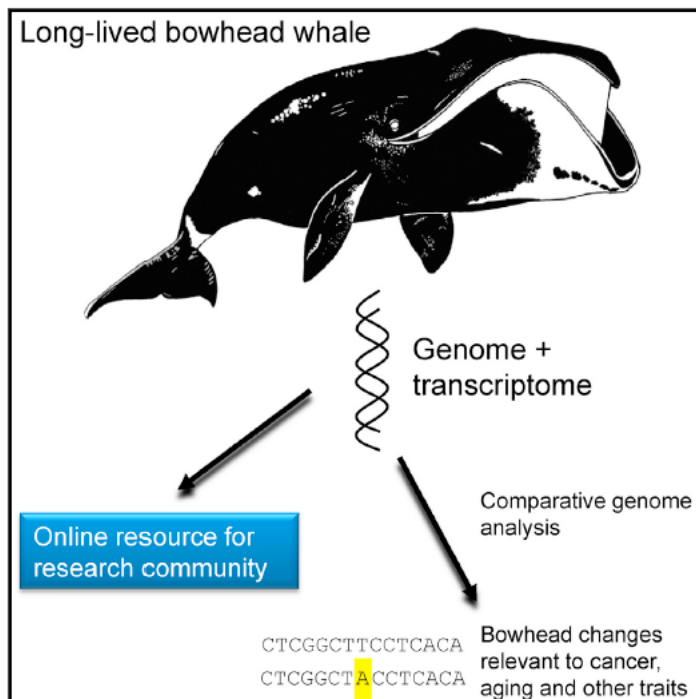
## 2. Annotation and analysis of the bowhead whale genome

### Introduction

The lifespan of some animals, including quahogs, tortoises, and certain whale species, is far greater than that of humans (Austad, 2010; Finch, 1990). It is remarkable that a warm-blooded species such as the bowhead whale (*Balaena mysticetus*) has not only been estimated to live over 200 years (the estimated age of one specimen was 211 with SE of 35 years), suggesting that it is the longest-lived mammal, but also exhibits very low disease incidence until an advanced age compared to humans (George et al., 1999; Philo et al., 1993). As in humans, the evolution of longevity in whales was accompanied by low fecundity and longer developmental time (Tacutu et al., 2013), as predicted by evolutionary theory. The cellular, molecular, and genetic mechanisms underlying longevity and resistance to age-related diseases in bowhead whales are unknown, but it is clear that in order to live so long, these animals must possess preventative mechanisms against cancer, immunosenescence and neurodegenerative, cardiovascular and metabolic diseases. In the context of cancer, whales, and bowhead whales in particular, must possess effective antitumor mechanisms. Indeed, given their large size (in extreme cases adult bowhead whales can weigh up to 100 tons and are therefore among the largest whales) and exceptional longevity, bowhead whale cells must have a significantly lower probability of neoplastic transformation relative to humans (Caulin and Maley, 2011; de Magalhães, 2013). Therefore, studying species such as bowhead whales that have greater natural longevity and resistance to age-related diseases than humans may lead to insights on the fundamental mechanisms of disease-resistance and aging.

As such, the results presented here are based on the annotation and genome-wide analysis of molecular evolution in protein-coding sequences of the bowhead whale following the sequencing and assembly of the genome (Keane et al., 2015).

Figure 2.1: workflow of the bowhead whale genome sequencing project (Keane et al., 2015).



My contribution to this publication included assessing the completeness of the assembly, identification of orthologs, calculation of dN/dS values, analysis of positive selection, identification of unique amino acid residues, manual annotation of genes, formatting of the data and results to be made available on the web portal and writing of the manuscript.

Gene annotation is acknowledged to be a challenging process and at the time of publication, it was appreciated that there were data-quality concerns with the results, relating in particular to the annotation. As a result, I subsequently undertook to re-annotate the bowhead genome in an attempt to generate an improved gene annotation, the results of which are also described herein.

## Methods

Completeness of the assembly was assayed with CEGMA which screens an assembly against a set of 248 core eukaryotic genes that are present in a wide range of taxa (Parra et al., 2007).

Putative genes were located in the assembly using the MAKER2 (Holt and Yandell, 2011) automated annotation pipeline, which used comparative and *de novo* prediction methods including BLASTX, Exonerate, SNAP, Genemark and Augustus. In addition to the cow, dolphin and human proteomes, two transcriptome assemblies, generated from specimens from Greenland and Alaska, were used as inputs to the comparative methods. Although an organism-specific repeat library should ideally be used also, unfortunately this was not available so the default provided with MAKER2 was instead employed. Repetitive elements were found with RepeatMasker (<http://www.repeatmasker.org/>). The annotation described in the published manuscript was generated using a single MAKER2 iteration. However, the recommended procedure is to employ the RNA-seq and protein data for the initial run and then use the output to iteratively train a gene predictor such as SNAP. Therefore I subsequently generated another annotation using a total of three iterative runs of MAKER2. The parameters used differed significantly from those employed in generating the published annotation. For the initial run, SNAP was disabled and the parameters to infer gene predictions directly from the RNA-seq and protein data (est2genome and protein2genome, respectively) were set. For the subsequent iterations, a SNAP HMM training file was generated and SNAP was enabled, with est2genome and protein2genome disabled. The commands and parameters used to generate the SNAP HMM training file were as specified in the MAKER tutorial ([http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/MAKER\\_Tutorial](http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/MAKER_Tutorial)).

To allow the identification of orthologous relationships with bowhead proteins, all cow protein sequences were downloaded from Ensembl (Flicek et al., 2013). Cow was used because it is the closest relative to the bowhead with a high-quality annotated genome available. First, BLASTp was used to find the best hit in the cow proteome for every predicted bowhead protein, and then the reciprocal best hit for each cow protein was defined as an ortholog.

To facilitate further studies of these animals, an online genome portal was constructed: The Bowhead Whale Genome Resource (<http://www.bowhead-whale.org/>). Its database structure, interface and functionality were adapted from our existing Naked Mole Rat Genome Resource (Keane et al., 2014). All data and results are available from the portal and supplemental methods and data files are also available on GitHub (<https://github.com/maglab/bowhead-whale-supplementary/>). My contribution to this resource was in relation to the preparation and formatting of the data and results that have been made available.

The CodeML program from the PAML package was used to calculate pairwise dN/dS ratios (Yang, 2007) using the ratio of nonsynonymous substitutions per nonsynonymous site (dN) to synonymous substitutions per synonymous site (dS), dN/dS (Yang, 2007). Specifically, these pairwise dN/dS ratios were calculated for bowhead coding sequences and orthologous sequences from minke, cow and dolphin, excluding coding sequences that were less than 50% of the length of the orthologous sequence. The results were then ranked by decreasing dN/dS and are available on the Bowhead Whale Genome Resource. In addition, the ratio of the bowhead-minke dN/dS value to the higher of the dN/dS values for minke-cow and minke-dolphin was calculated to identify genes that evolved more rapidly on the bowhead lineage.

An in-house Perl pipeline was used to align each bowhead protein with orthologs from nine other mammals: human (*Homo sapiens*), dog (*Canis familiaris*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), minke whale (*Balaenoptera acutorostrata*), cow (*Bos taurus*), dolphin (*Tursiops truncatus*), horse (*Equus caballus*) and elephant (*Loxodonta africana*), and then identify the unique bowhead amino acid residues. Gaps were excluded from the analysis, and a maximum of one unknown residue was allowed in species other than the bowhead. The results were ranked by the number of unique residues normalized by the protein length.

Manual annotation was attempted for a number of bowhead genes. In each case, the relevant bowhead scaffold was aligned with the orthologous cow sequence in addition to the human and mouse consensus coding sequences (Pruitt et al., 2009). Novel bowhead sequence was filtered unless supported by at least one transcript from the bowhead RNA-seq data.

## Results

The full and partial completeness of the bowhead whale draft genome assembly was evaluated as 93.15% and 97.18%, respectively, by the CEGMA pipeline (Parra et al., 2007), which is comparable to the minke whale genome assembly (Yim et al., 2014). Full completeness requires an alignment at least 70% of the protein length and partial completeness requires that a pre-computed minimum alignment score is exceeded.

The published annotation generated by MAKER2 contains 22,672 predicted protein-coding genes with an average length of 417 (median 307) amino acid residues. Orthologs with cow, human and mouse genes/proteins were identified based on similarity which allowed predicted gene symbols to be assigned to 15,831 bowhead genes.

As an initial assessment of coding genes that could be responsible for bowhead whale adaptations, bowhead coding sequences were used to calculate pairwise dN/dS ratios for 9,682, 12,685 and 11,158 orthologous coding sequences from minke whale (*Balaenoptera acutorostrata*), cow (*Bos taurus*) and dolphin (*Tursiops truncatus*), respectively. There are high levels of sequence conservation in the protein coding regions between bowhead and these species: 96% (minke), 92% (dolphin) and 91% (cow). This is unsurprising, however, given the long generation time of cetaceans and of the bowhead whale, in particular, with animals only reaching sexual maturity at >20 years (Tacutu et al., 2013). Because the minke whale is the closest relative to the bowhead (divergence time 25–30 million years ago) (Gatesy et al., 2013) with a sequenced genome and is smaller (<10 tons) and probably much shorter lived (maximum lifespan ~ 50 years) (Tacutu et al., 2013), comparisons between the bowhead and minke whale genomes may provide insights on the evolution of bowhead traits and of longevity, in particular. A number of aging and cancer-associated genes were observed among the 420 predicted bowhead-minke orthologs with dN/dS exceeding 1, including *suppressor of cytokine signaling 2* (SOCS2), *aprataxin* (APTX), *noggin* (NOG), and *leptin* (LEP). In addition, the top 5% genes with high dN/dS values for bowhead-minke relative to the values for minke-cow and minke-dolphin orthologs included *forkhead box O3* (FOXO3), *excision repair cross-complementing rodent repair deficiency, complementation group 3* (ERCC3), and *fibroblast growth factor receptor 1* (FGFR1). The data on dN/dS ratios are available on the genome portal to allow other researchers to do their own analysis and quickly retrieve gene(s) of interest.

In addition to codon-based models of evolution, bowhead whale specific amino acid substitutions were also identified. Specifically, orthologous sequences between the bowhead whale and nine other mammals were generated - a total of 4,358 alignments. Lineage-specific residues identified in this way have previously been shown to be indicative of significant changes in protein function (Tian



et al., 2013). This analysis revealed several proteins associated with aging and cancer among the top 5% of unique bowhead residues by concentration (i.e. normalized by protein length), including ERCC1 (*excision repair cross-complementing rodent repair deficiency, complementation group 1*), HDAC1 (*histone deacetylase 1*) and HDAC2 (Figure 2.2).

Figure 2.2: Partial alignment of bowhead *HDAC2* with mammalian orthologs. Unique bowhead residues are highlighted at human positions 68, 95, and 133.

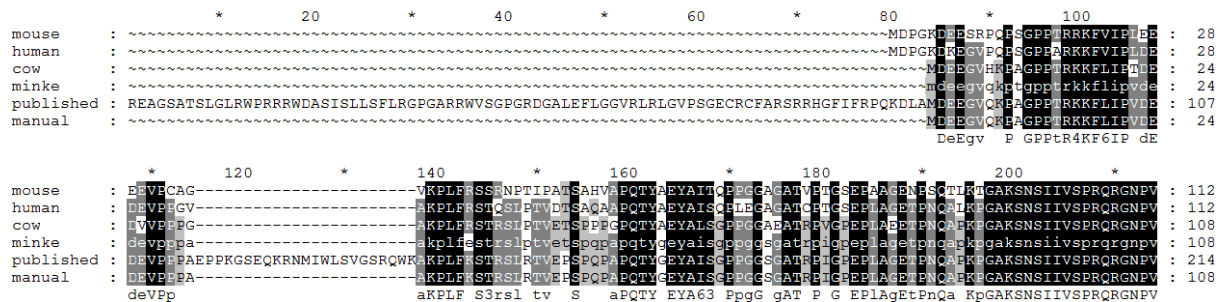
		*	180	*	200	*	220	*	240																																																																										
Cow	:	K	M	H	S	D	E	Y	I	K	F	L	R	S	I	R	P	D	N	M	S	E	Y	S	K	Q	M	Q	R	E	N	V	G	E	D	C	P	V	F	D	G	L	F	E	F	C	Q	L	S	T	G	G	S	V	A	G	V	K	L	N	R	Q	Q	T	D	M	A	V	N	W	A	G	G	L	H	H	A	K	K	S	E	:	147
Bowhead	:	K	M	H	S	D	E	Y	I	K	F	L	R	S	I	R	P	D	N	M	S	E	Y	S	K	Q	M	Q	R	S	N	V	G	E	D	C	P	V	F	D	G	L	F	E	F	C	Q	L	S	T	G	G	S	V	A	G	V	K	L	N	R	Q	Q	T	D	M	S	V	N	W	A	G	G	L	H	H	A	K	K	S	E	:	147
Rat	:	K	M	H	S	D	E	Y	I	K	F	L	R	S	I	R	P	D	N	M	S	E	Y	S	K	Q	M	Q	R	E	N	V	G	E	D	C	P	V	F	D	G	L	F	E	F	C	Q	L	S	T	G	G	S	V	A	G	V	K	L	N	R	Q	Q	T	D	M	A	V	N	W	A	G	G	L	H	H	A	K	K	S	E	:	147
Elephant	:	K	M	H	S	D	E	Y	I	K	F	L	R	S	I	R	P	D	N	M	S	E	Y	S	K	Q	M	Q	R	E	N	V	G	E	D	C	P	V	F	D	G	L	F	E	F	C	Q	L	S	T	G	G	S	V	A	G	V	K	L	N	R	Q	Q	T	D	M	A	V	N	W	A	G	G	L	H	H	A	K	K	S	E	:	147
Dolphin	:	K	M	H	S	D	E	Y	I	K	F	L	R	S	I	R	P	D	N	M	S	E	Y	S	K	Q	M	Q	R	E	N	V	G	E	D	C	P	V	F	D	G	L	F	E	F	C	Q	L	S	T	G	G	S	V	A	G	V	K	L	N	R	Q	Q	T	D	M	A	V	N	W	A	G	G	L	H	H	A	K	K	S	E	:	243
Dog	:	K	M	H	S	D	E	Y	I	K	F	L	R	S	I	R	P	D	N	M	S	E	Y	S	K	Q	M	Q	R	E	N	V	G	E	D	C	P	V	F	D	G	L	F	E	F	C	Q	L	S	T	G	G	S	V	A	G	V	K	L	N	R	Q	Q	T	D	M	A	V	N	W	A	G	G	L	H	H	A	K	K	S	E	:	147
Mouse	:	K	M	H	S	D	E	Y	I	K	F	L	R	S	I	R	P	D	N	M	S	E	Y	S	K	Q	M	Q	R	E	N	V	G	E	D	C	P	V	F	D	G	L	F	E	F	C	Q	L	S	T	G	G	S	V	A	G	V	K	L	N	R	Q	Q	T	D	M	A	V	N	W	A	G	G	L	H	H	A	K	K	S	E	:	147
Horse	:	K	M	H	S	D	E	Y	I	K	F	L	R	S	I	R	P	D	N	M	S	E	Y	S	K	Q	M	Q	R	E	N	V	G	E	D	C	P	V	F	D	G	L	F	E	F	C	Q	L	S	T	G	G	S	V	A	G	V	K	L	N	R	Q	Q	T	D	M	A	V	N	W	A	G	G	L	H	H	A	K	K	S	E	:	117
Minke	:	K	M	H	S	D	E	Y	I	K	F	L	R	S	I	R	P	D	N	M	S	E	Y	S	K	Q	M	Q	R	E	N	V	G	E	D	C	P	V	F	D	G	L	F	E	F	C	Q	L	S	T	G	G	S	V	A	G	V	K	L	N	R	Q	Q	T	D	M	A	V	N	W	A	G	G	L	H	H	A	K	K	S	E	:	117
Human	:	K	M	H	S	D	E	Y	I	K	F	L	R	S	I	R	P	D	N	M	S	E	Y	S	K	Q	M	Q	R	E	N	V	G	E	D	C	P	V	F	D	G	L	F	E	F	C	Q	L	S	T	G	G	S	V	A	G	V	K	L	N	R	Q	Q	T	D	M	A	V	N	W	A	G	G	L	H	H	A	K	K	S	E	:	147

Histone deacetylases play an important role in the regulation of chromatin structure and transcription (Lee et al., 1993) and have been associated with longevity in *Drosophila* (Rogina et al., 2002). ERCC1 is a member of the nucleotide excision repair pathway (Gillet and Scharer, 2006), and disruption results in greatly reduced lifespan in mice and accelerated aging (Weeda et al., 1997). As such, these represent candidates for involvement in adaptive genetic changes conferring disease resistance in the bowhead whale.

In addition to genes related to longevity, several interesting candidate genes emerged from the analysis of lineage-specific residues of potential relevance to other bowhead traits. Of note, a number of proteins related to sensory perception of sound were also identified with bowhead-specific mutations, including *otoraplin* (OTOR) and *cholinergic receptor, nicotinic, alpha 10* (CHRNA10), which could be relevant in the context of the bowhead's ability to produce high- and low-frequency tones simultaneously (Tervo et al., 2011). In addition, many proteins must play roles in the large differences in size and development between the bowhead and related species and the results reveal possible candidates for further functional studies; for example, in the top ten proteins, SNX3 (*sorting nexin 3*) has been associated in one patient with eye formation defects and microcephaly (Vervoort et al., 2002) and WDR5 (*WD repeat-containing protein 5*) has been associated with osteoblast differentiation and bone development (Gori et al., 2006).

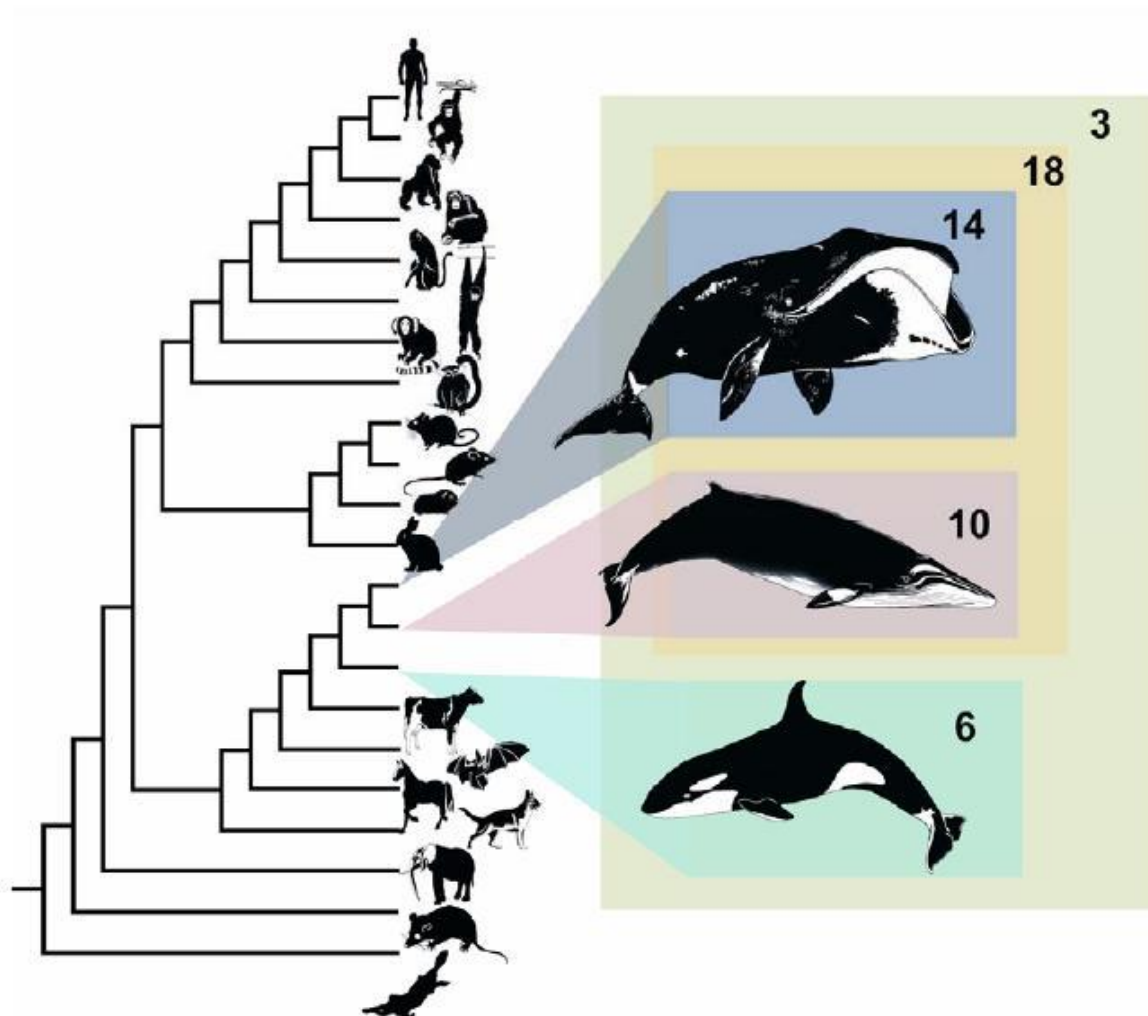
The results described thus far were based on the published gene annotation which was generated using a single iteration of MAKER2, however there were indications that the quality of this annotation was questionable. Specifically, visual inspection of a number of the alignments generated for genes initially identified as positively selected revealed some probable annotation errors, which were subsequently confirmed following manual annotation incorporating the transcriptome data (see Methods). For example, the ERCC1 gene was incorrectly annotated as starting upstream of orthologs in other species in addition to containing an erroneous read-through first intron (Figure 2.3).

Figure 2.3: Partial alignment of the ERCC1 protein in mouse, human, cow, minke and bowhead (published and manual annotations).



The level of annotation errors observed necessitated the introduction of stringent filtering of candidate genes in order to remove spurious results. As a direct consequence, ultimately only 14 bowhead and 10 minke genes were found to exhibit evidence of positive selection using codon-based models of evolution (Figure 2.4), which is far fewer than typically reported in genome-wide analyses.

Figure 2.4: Phylogeny of mammals used in comparison of selective pressure variation (Keane et al., 2015). The number of candidate genes under positive selection on each lineage is indicated.



By contrast, the analysis of the minke whale genome reported 279 positively selected genes by comparison with cow and pig using the branch-site likelihood ratio test (Yim et al., 2014).

In an attempt to address these annotation issues, subsequent to publication I generated another gene annotation using multiple iterations of MAKER2 (see Methods). As an initial assessment of annotation quality, I examined the *Hox* gene cluster as it is well-conserved and characterised in other vertebrates (Zákány et al., 2004). In addition, several of these genes (HOXA5, HOXB1, HOXB2, HOXB5, HOXD12 and HOXD13) were reported to have been positively selected on the whale lineage compared to terrestrial mammals in the analysis of the minke whale genome (Yim et al., 2014). As

such, I used the protein lengths (Table 2.1) and alignments (Figures 2.5 – 2.10) of these genes in order to assess the quality of the annotations generated.

Table 2.1: proteins lengths of six annotated *Hox* genes in human, mouse, cow, minke and bowhead. The results of three bowhead annotations are included: the published annotation and also both the initial and final iterations of the MAKER2 annotation which I subsequently generated.

	Protein length						
	Human	Mouse	Cow	Minke	Published	Initial	Final
HOXA5	270aa	270aa	270aa	381aa	661aa	270aa	1270aa
HOXB1	301aa	297aa	302aa	298aa	299aa	298aa	335aa
HOXB2	356aa	354aa	354aa	?	787aa	492aa	751aa
HOXB5	269aa	269aa	251aa	269aa	269aa	269aa	181aa
HOXD12	270aa	268aa	271aa	270aa	321aa	270aa	1530aa
HOXD13	343aa	339aa	327aa	264aa	205aa	204aa	1530aa

Figure 2.5: Partial alignment of the HOXA5 protein in human, mouse, cow, minke and bowhead (published, initial and final annotations).

```

          960          *          980          *          1000          *          1020          *          1040          *          1060
human      : ~~~~~MSSYFVNSF : 9
mouse      : ~~~~~MSSYFVNSF : 9
cow        : ~~~~~MSSYFVNSF : 9
minke      : rrrvagtrlgrarhssrlhptpplvpeftsrghqagfttgqqkhvirsrtpylgayvggnqvhvpvisiihhlckgaid-agttashkssthikkqmssyfvnsf : 120
published  : AEAKAGEDQARAATSLGLPPCLLLLPNFSSLLLPSSGGLS-----QLQ-G-----SPALCKRLSIYFLQEKKKKKNLTHSQSHSYFVNSF : 399
initial    : ~~~~~MSSYFVNSF : 9
final      : AEAKAG-----DSYFVNSF : 968
                      SYFVNSF

          *          1080          *          1100          *          1120          *          1140          *          1160
human      : CGRYPNGPDYQLHNYGDHSSVSEQFRDSASMHSGRYGYGYNGMDLSVGRSGSGHFGSGGERARSYAASASAAPAEPRYSQPATSTHSPQPDPLPCSAVAPSPGSDSH : 115
mouse      : CGRYPNGPDYQLHNYGDHSSVSEQFRDSASMHSGRYGYGYNGMDLSVGRSGSGHFGSGGERARSYAAGASAAAPAEPRYSQPATSTHSPQPDPLPCSAVAPSPGSDSH : 115
cow        : CGRYPNGPDYQLHNYGDHSSVSEQFRDSASMHSGRYGYGYNGMDLSVGRSGSGHFGSGGERARSYAAGASAAAPAEPRYSQPATSTHSPQPDPLPCSAVAPSPGSDSH : 115
minke      : cgrypngpdyqlhnygdhssvseqfrdsasmhsgrgygyngmdlsvgrsgsgghfgsggerarsyaasasaapaeprysqpatsthspqpdplpcsavapspgtdsh : 226
published  : CGRYPNGPDYQLHNYGDHSSVSEQFRDSASMHSGRYGYGYNGMDLSVGRSGSGHFGSGGERARSYAASAGAAPAEPRYSQPATSTHSPQPDPLPCSAVAPSPGSDSH : 505
initial    : CGRYPNGPDYQLHNYGDHSSVSEQFRDSASMHSGRYGYGYNGMDLSVGRSGSGHFGSGGERARSYAASAGAAPAEPRYSQPATSTHSPQPDPLPCSAVAPSPGSDSH : 115
final      : CGRYPNGPDYQLHNYGDHSSVSEQFRDSASMHSGRYGYGYNGMDLSVGRSGSGHFGSGGERARSYAASAGAAPAEPRYSQPATSTHSPQPDPLPCSAVAPSPGSDSH : 1074
                      CGRYPNGPDYQLHNYGDHSSVSEQFRDSASMHSGRYGYGYNGMDLSVGRSGSGHFGSGGERARSYAA A AAPAEPRYSQPATSTHSPQPDPLPCSAVAPSPG3DSH

          *          1180          *          1200          *          1220          *          1240          *          1260          *
human      : HGGKNSISNSSGASADAGSTHISREGVGTASGAEDAPASSEQASAQSEPSAPPAPQPIYPWMRKLHISH----- : 187
mouse      : HGGKNSLGNSSGASANAGSTHISREGVGTASGAEDAPASSEQASAQSEPSAPPAPQPIYPWMRKLHISH----- : 187
cow        : HGGKNSLGNSSGASANAGSTHISREGVGTASGAEDAPASSEQASAQSEPSAPPAPQPIYPWMRKLHISH----- : 187
minke      : hggknslnssgasantgsthissregvgtasgaedapasseqasaqsepsappapqpiypwmrklhish----- : 298
published  : HGGKNSLGNSSGASANAGSTHISREGVGTASGAEDAPASSEQASAQSEPSAPPAPQPIYPWMRKLHISH----- : 577
initial    : HGGKNSLGNSSGASANAGSTHISREGVGTASGAEDAPASSEQASAQSEPSAPPAPQPIYPWMRKLHISH----- : 187
final      : HGGKNSLGNSSGASANAGSTHISREGVGTASGAEDAPASSEQASAQSEPSAPPAPQPIYPWMRKLHISHGKEIREGWAGERGGRLVELFSPVQDSGWSSWRTL : 1180
                      HGGKNSLgNSSGASAlaGSTHISREGVGTASgAEEDAPASSEQAsAQSEPSAPPAPQPIYPWMRKLHISH

```

Figure 2.6: Full alignment of the HOXB1 protein in human, mouse, cow, minke and bowhead (published, initial and final annotations).

```

      *      20      *      40      *      60      *      80      *      100
human   : MDYNRMNSFLEYPLCNRPSPAYS AHSAPTSTFPSSAQVDSYASEGRYGGGLSSPAFQQNSGYPAQCPPSTLGVFPFPSSAFSGYAPAACSPSYGPSQYYPLGQSEGD : 107
mouse   : MDYNRMNSFLEYPLCNRPSPAYS ---APTSTFPSCSAFVDSYAGESRYGGGLPSSALQQNSGYFVQCPPSSLGVSFPPSAFSGYAPAACNSYGPSQYYSVGQSEGD : 104
cow      : MDYNRMNSFLEYPLCNRPSPAYS AHSAPTSTFPSSASAVDSYAGETRYGGGLPSSALQQNSGYLAQCPPSLGVFPFPSSATSGYSPAACSA SYGSSQYYPLGQLEGD : 107
minke    : mdynrmnsfleyplcnrapsays ---aptsfppssapsvdsyagetryggglpsalqqnsgypachppsalgvfpfpssatsgyapaacspsygpsqyyplgqlegd : 104
published : MDYNRMNSFLEYPLCNRAPSPAYS ---APTSTFPSSAPSVDSYAGETRYGGGLPSPALQQNSGYPAQHPPSALGVFPFPSSATSGYAPAACSPSYGPSQYYPLGQFEGD : 104
initial  : MDYNRMNSFLEYPLCNRAPSPAYS ---APTSTFPSSAPSVDSYAGETRYGGGLPSPALQQNSGYPAQHPPSALGVFPFPSSATSGYAPAACSPSYGPSQYYPLGQFEGD : 104
final    : MDYNRMNSFLEYPLCNRAPSPAYS ---APTSTFPSSAPSVDSYAGETRYGGGLPSPALQQNSGYPAQHPPSALGVFPFPSSATSGYAPAACSPSYGPSQYYPLGQFEGD : 104
          MDYNRMNSFLEYPLCNRPSPAYS APTSTFPSSA VDSYAG E RYGGGLPs AlQQNSGYpaQ PPS LGVpFPSSa SGYaPAACspSYGpsQYYp6GQ EGD

      *      120      *      140      *      160      *      180      *      200      *
human   : GGYFHPSSYGAQLGGLSDGYGAGGAGGPGYPPEQHPPYGNEQTASFAFAYADLLSEDKETPCPSEPNTPTARTFDWMKVKNPPKT----- : 192
mouse   : GSYFHPSSYGAQLGGLSDSYGAGGVGSGPYPPPQPPYGTEQTATFASAYDLLSEDKESPCSSEPSTITERTFDWMKVKNPPKT----- : 188
cow      : GGYFHPSSYGAQLGGLSDGYGTAGVGPGYPPPPQPPYGNEQTGSFAPACADLLSEDKESACTSETSTPTACTFDWMKVKNPPKT----- : 193
minke    : ggyfhpssygaqlgal-sdgygtagvgpgyppppppygneqtnfapayaellseakespcpsetslptartfdwmkvkrnppkt----- : 189
published : GGYFHPSSYGAQLGGLSDGYGTAGVGPGYPPPPQPPYGNEQTGNFAPAYAELLSEAKESPCPSETSLPTARTFDWMKVKNPPKT----- : 189
initial  : GGYFHPSSYGAQLGGLSDGYGTAGVGPGYPPPPQPPYGNEQTGNFAPAYAELLSEAKESPCPSETSLPTARTFDWMKVKNPPKT----- : 189
final    : GGYFHPSSYGAQLGGLSDGYGTAGVGPGYPPPPQPPYGNEQTGNFAPAYAELLSEAKESPCPSETSLPTARTFDWMKVKNPPKTGRAQTWPPPLWGSERLCCFCT : 210
          GgyFHPSSYGAQLG L sDgYG G GpGPYPpPqPpYGnEQT FApAya LLSE KE3pC SE s pTarTFDWMKVKNPPKT

      220      *      240      *      260      *      280      *      300      *      320
human   : -----AKVSEFGLGSEFSGLRTNFTTRQLTELEKEFHFNKYLSRARRVEIAATLELNETQVKIWFQNRRMKQKKREREGGRVPPAPPGCPKEAAGDA : 283
mouse   : -----AKVSELGLGAPGGLRTNFTTRQLTELEKEFHFNKYLSRARRVEIAATLELNETQVKIWFQNRRMKQKKREREGGRMEAGPPGCPKEAAGDA : 279
cow      : -----AKVSELGLGAPGGLRTNFTTRQLTELEKEFHFNKYLSRARRVEIAATLELNETQVKIWFQNRRMKQKKREREGGRVPPAPPGCPKEATGDT : 284
minke    : -----akvselglgapgglrtnfttrqltelekefhfnkylsrarrveiaatlelnetqvkifqnrrmkqkkregeggrvppapsgcpkeaagda : 280
published : -----AKVSELGLGAPGGLRTNFTTRQLTELEKEFHFNKYLSRARRVEIAATLELNETQVKIWFQNRRMKQKKREREGGRVPPAPSGCPKEAAGDA : 280
initial  : -----AKVSELGLGAPGGLRTNFTTRQLTELEKEFHFNKYLSRARRVEIAATLELNETQVKIWFQNRRMKQKKREREGGRVPPAPSGCPKEAAGDA : 280
final    : EVSWAGFLSLRRRVEAKVSELGLGAPGGLRTNFTTRQLTELEKEFHFNKYLSRARRVEIAATLELNETQVKIWFQNRRMKQKKREREGGRVPPAPSGCPKEAAGDA : 317
          AKVSELGLGaPgGLRTNFTTRQLTELEKEFHFNKYLSRARRVEIAATLELNETQVKIWFQNRRMKQKKREREGGR6PpaP GCPKEAaGDa

```



Figure 2.7: Partial alignment of the HOXB2 protein in human, mouse, cow and bowhead (published, initial and final annotations).

```

*      340      *      360      *      380      *      400      *      420
human   : ~~~~~~MNFEFEREIGFINS : 14
mouse   : ~~~~~~MNFEFEREIGFINS : 14
cow      : ~~~~~~MNFEFEREIGFINS : 14
published : HVLQANGGAYGTPTMQGSPVYVGGGGYADPLPPPAGPSLYGLNHLSHHPSGNLDYNGAPPMAPSQHGGPCDPHPTYTDLSSHHAPPPQGRIQEAPKLTHLEIGFINS : 427
initial   : HVLQANGGAYGTPTMQGSPVYVGGGGYADPLPPPAGPSLYGLNHLSHHPSGNLDYNGAPPMAPSQHGGPCDPHPTYTDLSSHHAPPP----- : 264
final     : HVLQANGGAYGTPTMQGSPVYVGGGGYADPLPPPAGPSLYGLNHLSHHPSGNLDYNGAPPMAPSQHGGPCDPHPTYTDLSSHHAPPPQAMNFE---FEREIGFINS : 392
                                         eigfins

*      440      *      460      *      480      *      500      *      520      *
human   : QPSLAECLTSFPAVLETFQTSSIKESTLIPIPPPPFEQTFPSLQPGASTLQRPRSQKRAEDGPAIPPPPPPELPAAPPAPEFPWMKEKKS AKKPSQSATSPSPAASAV : 121
mouse   : QPSLAECLTSFPAVLETFQTSSIKESTLIPIPPPPFEQTFPSLQPGASTLQRPGSQKQAGDGPALRSP--PELPVAPPAPPEFPWMKEKKS TKKPSQSAASPSPAASSV : 119
cow      : QPSLAECLTSFPAVLETFQTSSIKESTLIPIPPPPFEQTFPSLQPGASTLQRPGSQKRAEDGPAIPPL--PELPAAPLAPEFPWMKEKKS AKKPSQSAASP-PAASSV : 118
published : QPSLAECLTSFPAVLETFQTSSIKESTLIPIPPPPFEQTFPSLQPGASTLQRPGSQKRAEDGPAIPPP---PEFPAAPLAPQFPWMKEKKS AKKPSQSATSP-PAASSV : 530
initial   : ----- : -
final     : QPSLAECLTSFPAVLETFQTSSIKESTLIPIPPPPFEQTFPSLQPGASTLQRPGSQKRAEDGPAIPPP---PEFPAAPLAPQFPWMKEKKS AKKPSQSATSP-PAASSV : 495
          qpslaeccltsfpavletfqtssikestlippppp eqtfpslq gastlqrp sqk a dgpai pp p p ap ap fpwmkekks kkpsqsa sp paas v

*      540      *      560      *      580      *      600      *      620      *      640
human   : PASGVGSP-----ADGGLGPEAGGGGARRLRTAYTNTQLLELEKEFHFNKYLCRPRRVEIAALLDLTERQVKVWFQNRMMKHKRQTQHREPP : 208
mouse   : PASGVGSP-----SDGPGLEPCGGSGSRRRLRTAYTNTQLLELEKEFHFNKYLCRPRRVEIAALLDLTERQVKVWFQNRMMKHKRQTQHREPP : 206
cow      : LASGVGSP-----ADGPGLEPAAGGGARRLRTAYTNTQLLELEKEFHFNKYLCRPRRVEIAALLDLTERQVKVWFQNRMMKHKRQTQHREPP : 205
published : PASGVGSPAGRRGWAKDTARVDAQPAFPAADGPGLEPAAGGGARRLRTAYTNTQLLELEKEFHFNKYLCRPRRVEIAALLDLTERQVKVWFQNRMMKHKRQTQHREPP : 637
initial   : -----CDGPGLEPAAGGGARRLRTAYTNTQLLELEKEFHFNKYLCRPRRVEIAALLDLTERQVKVWFQNRMMKHKRQTQHREPP : 343
final     : PASGVGSPAGRRGWAKDTARVDAQPAFPAADGPGLEPAAGGGARRLRTAYTNTQLLELEKEFHFNKYLCRPRRVEIAALLDLTERQVKVWFQNRMMKHKRQTQHREPP : 602
          as vgs p DGpGL Ea GgGaRRRLRTAYTNTQLLELEKEFHFNKYLCRPRRVEIAALLDLTERQVKVWFQNRMMKHKRQTQHREPP

```

Figure 2.8: Full alignment of the HOXB5 protein in human, mouse, cow, minke and bowhead (published, initial and final annotations).

```

      *      20      *      40      *      60      *      80      *      100
human   : MSSYFVNSFSGRYPNGPDYQLLNYGSGSSLSGSYRDPAAAMHTGSYGYNNGMDLSVNRSSASSSHFGAVGESSRAFPAPAQEPRFRQAASSCSLSSPESLPCTNGDSHG : 109
mouse   : MSSYFVNSFSGRYPNGPDYQLLNYGSGSSLSGSYRDPAAAMHTGSYGYNNGMDLSVNRSSASSSHFGAVGESSRAFPAPSAQEPRFRQATSSCSLSSPESLPCTNGDSHG : 109
cow      : MSSYFVNSFSGRYPNGPDYQLLNYGSGSSLSGSYRDPAAAMHTGSYGYNNGMDLSVNRSSASSSHFGAVGESSRAFPAPAQEPRFRQAASSCSLSSPESLPCTNGDSHG : 109
minke    : mssyfvnsfsgrypnngpdyqllnygsqsslsqsyrdpaamhtgsygyngmdlsvnrssassshfgavgessrafsapsqeprfrqaasscsllsspeslpctngdshg : 109
published : MSSYFVNSFSGRYPNGPDYQLLNYGSGSSLSGSYRDPAAAMHTGSYGYNNGMDLSVNRSSASSSHFGAVGESSRAFPAPSQEPRFRQAASSCSLSSPESLPCTNGDSHG : 109
initial  : MSSYFVNSFSGRYPNGPDYQLLNYGSGSSLSGSYRDPAAAMHTGSYGYNNGMDLSVNRSSASSSHFGAVGESSRAFPAPSQEPRFRQAASSCSLSSPESLPCTNGDSHG : 109
final    : ~~~~~~MHTGSYGYNNGMDLSVNRSSASSSHFGAVGESSRAFPAPSQEPRFRQAASSCSLSSPESLPCTNGDSHG : 70
          mssyfvnsfsgrypnngpdyqllnygsqsslsqsyrdpaamhtgsygyngmdlsvnrssassshfgavgessrafpAp QEPRFRQAaSSCSLSSPESLPCTNGDSHG

      *      120      *      140      *      160      *      180      *      200      *      2
human   : AKPSASSPSDQATPASSSANFTEIDEASASSEPEEAASQLSSPSLARAQPEPMATSTAAPGQTPQIFPWMRKLHISHDMTGPDGKRARTAYTRY--QTLEIEKEEFHFN : 216
mouse   : AKPSASSPSDQATPASSSANFTEIDEASASSEPEEAASQLSSPSLARAQPEPMATSTAAPGQTPQIFPWMRKLHISHDMTGPDGKRARTAYTRY--QTLEIEKEEFHFN : 216
cow      : AKPSASSPSDQATPASSSANFTEIDEASASSEPEEAASQLSSPSLARAQPEPMATSTAAPGQTPQIFPWMRKLHISHDMTGPDGKRARTAYTRY--QTLEIEKEEFHFN : 216
minke    : akpsasspsdqatpasssanfteideasvssepeeaasqlsspslaragqepmatstaapegqtpqifpwmrklhishdmtgpdgkrartaytry--qtlelekeefhfn : 216
published : AKPSASSPSDQATPASSSANFTEIDEASVSSEPEEAASQLSSPSLARAQPEPMATSTAAPGQTPQIFPWMRKLHISHDMTGPDGKRARTAYTRY--QTLEIEKEEFHFN : 216
initial  : AKPSASSPSDQATPASSSANFTEIDEASVSSEPEEAASQLSSPSLARAQPEPMATSTAAPGQTPQIFPWMRKLHISHDMTGPDGKRARTAYTRY--QTLEIEKEEFHFN : 216
final    : AKPSASSPSDQATPASSSANFTEIDEASVSSEPEEAASQLSSPSLARAQPEPMATSTAAPGQTPQIFPWMRKLHISHGNSRSRSPFCLLGTPSPRLGGGRGLENLAFP : 179
          AKPSASSPSdQATpASSSANFTEIDEAS SSEPEEAASQLSSPSLARAQPEPMA STAAPGQTPQIFPWMRKLHISHdm3gpdgkrarta5try qtleleEke hFn

      20      *      240      *      260      *
human   : RYLTRRRRIEIAHALCLSERQIKIWFQNRMMKWKKNKLKMSMLATAGSAFQP : 269
mouse   : RYLTRRRRIEIAHALCLSERQIKIWFQNRMMKWKKNKLKMSMLATAGSAFQP : 269
cow      : RYLTRRRRIEIAHALCLSERQIKIWFQNRMMKWKK~~~~~ : 251
minke    : ryltrrrrieiahalclserqikiwfnrrmkwkkdnklksmlatagsafqp : 269
published : RYLTRRRRIEIAHALCLSERQIKIWFQNRMMKWKKNKLKMSMLATAGSAFQP : 269
initial  : RYLTRRRRIEIAHALCLSERQIKIWFQNRMMKWKKNKLKMSMLATAGSAFQP : 269
final    : HI~~~~~ : 181
          ryltrrrrieiahalclserqikiwfnrrmkwkk

```



Figure 2.9: Partial alignment of the HOXD12 protein in human, mouse, cow, minke and bowhead (published, initial and final annotations).

```

                220      *      240      *      260      *      280      *      300      *      3
human      : ~~~~~~MCERSLYRAGYVGSLLNLQSPDSFYFSNLRPNNGQLAALPISYPRGALPWATTPASCAP : 60
mouse      : ~~~~~~MCERSLYRAGYVGSLLNLQSPDSFYFSNLRPNNGQLAALPISYPRGALPWATTPASCAP : 60
cow        : ~~~~~~LEMCERSLYRAGYVGSLLNLQSPDSFYFSNLRPNNGQLAALPISYPRGALPWATTPASCAP : 62
minke      : ~~~~~~mcerslyragyvgsllnlqspdsfyfsnlrpnngqlaalpsisyprgalpwattpascap : 60
published  : ~~~~~~MCERSLYRAGYVGSLLNLQSPDSFYFSNLRPNNGQLAALPISYPRGALPWATTPASCAP : 60
initial    : ~~~~~~MCERSLYRAGYVGSLLNLQSPDSFYFSNLRPNNGQLAALPISYPRGALPWATTPASCAP : 60
final      : AINKFINKDKRRRISAATNLSEKQVTIWFQNRVKDKKIVSKLKDA~MCERSLYRAGYVGSLLNLQSPDSFYFSNLRPNNGQLAALPISYPRGALPWATTPASCAP : 318
                MCERSLYRAGYVGSLLNLQSPDSFYFSNLRPNNGQLAALPISYPRGALPWATTPASCAP

                20      *      340      *      360      *      380      *      400      *      420
human      : AQPAGATAFGGFSQPYLAGSGPLGLQPTAKDGEEDQAKFYAEAAAGPEERGRTRPSEAPESSLAPAAVAALKAAKYDYAGVGRATPGSTLLQGAPCAPGFKDDT : 166
mouse      : AQPATASAFGGFSQPYLTGSGPIGLQSPGAKDGEEDQVKFYTPDAPTASEERSRTRPPEAPESSLVH--SALKGTKYDYAGVGRATPGSATLLQGAPCASSFKEDT : 164
cow        : AQPAGATAFGGFSQPYLAGSGPLGLQPLGAKDGEEDQAKYYPDAAGPEERGRARPPFIPESSLAPAAAALKAAKYDYAGMGRVAPGSSALLEGTPCSAQFKDDA : 168
minke      : aqpasatnfgsfqpylagsgplglqplgakdgseeqakfytpdaagpeergrrarppfi pesslapaaaalkaakydyagvgrvapgssallegtpcssqfkdet : 166
published  : AQPAGATNFGSFSQPYLAGSGPLGLQPLGAKDGEEDQAKFYTPDAAAGPEERGRARPPFIPESSLAPAAAALKAAKYDYAGVGRVAPGSSALLEGTPCSSGFKDET : 166
initial    : AQPAGATNFGSFSQPYLAGSGPLGLQPLGAKDGEEDQAKFYTPDAAAGPEERGRARPPFIPESSLAPAAAALKAAKYDYAGVGRVAPGSSALLEGTPCSSGFKDET : 166
final      : AQPAGATNFGSFSQPYLAGSGPLGLQPLGAKDGEEDQAKFYTPDAAAGPEERGRARPPFIPESSLAPAAAALKAAKYDYAGVGRVAPGSSALLEGTPCSSGFKDET : 424
                AQPA A3 FG FSQPYLaGSGP6GLQp gAKDG EeQaK5YtPdAaa pEERgR RPpF PESSLapa aALKaaKYDYaG6GR aPGS LL2G PC gFKd t

                *      440      *      460      *      480      *      500      *      520      *
human      : KGPLNLMNTVQAGVASCLRP SLPD-----GLPWGAAPGRARKKRKP YTKQQIAELENEFI : 222
mouse      : KGPLNLMNTVQAGVASCLRP SLPD-----GLPWGAAPGRARKKRKP YTKQQIAELENEFI : 220
cow        : KGPLNLMNTVQAGVASCLRP SLPD-----GLPWGAAPGRARKKRKP YTKQQIAELENEFI : 223
minke      : kgplnlmmtvqagvasclrp slpd-----glpwgaapgrarkkrkpytkqqiaeleneffi : 222
published  : KGPLNLMNTVQAGVASCLRP SLPDGKGWPRSASDAPRAGVGGGVQGRVCRGREPPGAGRQPGAGRADWGWGCVA GLPWGAAPGRARKKRKP YTKQQIAELENEFI : 272
initial    : KGPLNLMNTVQAGVASCLRP SLPD-----GLPWGAAPGRARKKRKP YTKQQIAELENEFI : 222
final      : KGPLNLMNTVQAGVASCLRP SLPDGKGWPRSASDAPRAGVGGGVQGRVCRGREPPGAGRQPGAGRADWGWGCVA GLPWGAAPGRARKKRKP YTKQQIAELENEFI : 530
                KGPLNLMNT6Q AGVASCL pSLPD GLPWGAAPGRARKKRKP YTKQQIAELENEFI

```

Figure 2.10: Partial alignment of the HOXD13 protein in human, mouse, cow, minke and bowhead (published, initial and final annotations).

```

          *      20      *      40      *      60      *      80      *      100
human   : MSRAGSWDM DGLRADGGGAGGAPASSSSSSVAAAAASGQCRGFLSAPVFAGTHSGRAAAAAAAAAAAAAASGFAYPGTSERTGSSSSSSSSAVVAARPEAPPAKE : 106
mouse   : MSRSGTWDM DGLRADGGAAGAAPASSSSS----VAAPGQCRGFLSAPVFAGTHSGRAAAAAAAAA--AAAAASSFAYPGTSERTGSSSSSSSSAVIATRPEAPVAKE : 101
cow     : ~~~~~~ASGRSLTGAAA--RLSPQSSSSSVAAAAPGQCRGFLSAPVFAGTHSGRAAAAAAAAAAAAAASGFAYPGTSERAGSASSSSSSAVVAARPEAPSAKE : 96
minke   : ~~~~~~gtseragsassssssavvaarpeassske : 29
published : ~~~~~~ : -
initial  : ~~~~~~ : -
final    : ~~~~~~RAGSASSSSSSAVVAARPEASSAKE : 25

          *      120      *      140      *      160      *      180      *      200      *
human   : CPAPTPAAAAAPPAPALGYGYHFGNGYYSCRM SHGVLQQNALKSSPHASLGGFPVEKYMDVSGLASSSVPA NEVPARAKEVSFYQGYTSPYQHVPGYIDMVST : 212
mouse   : CPAPAAAATAAAPPAPALGYGYHFGNGYYSCRM SHGVLQQNALKSSPHASLGGFPVEKYMDVSGLASSSVPA NEVPARAKEVSFYQGYTSPYQHVPGYIDMVST : 207
cow     : CPAPGA--AAAAPPAPALGYGYHFGNGYYSCRM SHGVLQQNALKSSPHASLGGFPVEKYMDVSGLASSSVPA NEVPARAKEVSFYQGYTSPYQHVPGYIDMVST : 200
minke   : cpapg--aaaaappgapalgygyhfgngyyscrm shgvlqqnalkssphaslggfpvekymdvsglasssvpane vparakevsfyqgytspyqhvpgyidmvst : 133
published : ~~~~~~MSHGVLQQNALKSSPHASLGGFPVEKYMDVSGLASSSVPA NEVPARAKEVSFYQGYTSPYQHVPGYIDMVST : 73
initial  : ~~~~~~MSHGVLQQNALKSSPHASLGGFPVEKYMDVSGLASSSVPA NEVPARAKEVSFYQGYTSPYQHVPGYIDMVST : 73
final    : CPAPGA--AAAAPPAPALGYGYHFGNGYYSCRM SHGVLQQNALKSSPHASLGGFPVEKYMDVSGLASSSVPA NEVPARAKEVSFYQGYTSPYQHVPGYIDMVST : 129
          MSHGVLQQNALKSSPHASLGGFPVEKYMDVSGLASSSVPA NEVPARAKEVSFYQGYTSPYQHVPGYIDMVST

          220      *      240      *      260      *      280      *      300      *      3
human   : FGSGEPRHEAYISMEGYQSWTLANGWNSQVYCAKDQPGGSHFWKSSFP GDVALNQPDMCVYRRGRKKRVPYTKLQLKELENEYAINKFINKDKRRRISAATNLSE : 318
mouse   : FGSGEPRHEAYISMEGYQSWTLANGWNSQVYCAKDQPGGSHFWKSSFP GDVALNQPDMCVYRRGRKKRVPYTKLQLKELENEYAINKFINKDKRRRISAATNLSE : 313
cow     : FGSGEPRHEAYISMEGYQSWTLANGWNSQVYCAKDQPGGSHFWKSSFP GDVALNQPDMCVYRRGRKKRVPYTKLQLKELENEYAVNKFINKDKRRRISAATNLSE : 306
minke   : fgsgeprheayismegyqswtlangwnsqvycakdqpqgshfwkssfp gdvalnqpdmcvyrgrkrvp ytklqlkeleneyainkfinkdkrrrisaatnlse : 239
published : FGSGEPRHEAYISMEGYQSWTLANGWNSQVYCAKDQPGGSHFWKSSFP GDVALNQPDMCVYRRGRKKRVPYTKLQLKELENEYAINKFINKDKRRRISAATNLSE : 179
initial  : FGSGEPRHEAYISMEGYQSWTLANGWNSQVYCAKDQPGGSHFWKSSFP GDVALNQPDMCVYRRGRKKRVPYTKLQLKELENEYAINKFINKDKRRRISAATNLSE : 179
final    : FGSGEPRHEAYISMEGYQSWTLANGWNSQVYCAKDQPGGSHFWKSSFP GDVALNQPDMCVYRRGRKKRVPYTKLQLKELENEYA6NKFINKDKRRRISAATNLSE : 235
          FGSGEPRHEAYISMEGYQSWTLANGWNSQVYCAKDQPGGSHFWKSSFP GDVALNQPDMCVYRRGRKKRVPYTKLQLKELENEYA6NKFINKDKRRRISAATNLSE

```

From the results presented in Table 2.1, it is clear that the human and mouse proteins are all practically identical in length, which provides a useful indication of the expected protein length for the remaining annotations. Four of the six cow protein lengths are very similar to human and mouse, however HOXB5 and HOXD13 are significantly shorter indicating potential issues with these annotations. There appears to be more serious issues with the minke annotations as the lengths of the HOXA5 and HOXD13 proteins are very different. In addition, there is no available annotation for HOXB2, which raises a fundamental question in relation to whether it is possible to regard a prediction of positive selection for this gene with confidence.

The published bowhead annotation seems to have further serious issues, with only HOXB1 and HOXB5 not exhibiting large differences in protein length. Therefore it is encouraging that the initial annotation I subsequently generated appears to have resulted in an improvement. Specifically, four of the annotated proteins (HOXA5, HOXB1, HOXB5 and HOXD12) now have a length close to that which would be expected. However it is disappointing to note that the subsequent SNAP iterations, far from incrementally improving annotations, have apparently had exactly the opposite effect. In particular, a large amount of spurious sequence appears to have been added to several genes, particularly HOXA5, HOXB2, HOXD12 and HOXD13. Given that the results of the initial annotation look promising, this suggests that a confounding factor such as attempting to incorporate numerous splice variants into a single gene model may have contributed to the issues observed with the results of the SNAP iterations. The lack of a bowhead-specific repeat library could also provide an explanation for this issue. Many transposon open reading frames can look like true host genes to tools such as SNAP, resulting in portions being erroneously added as additional exons to gene annotations thereby corrupting the results. Therefore it would be extremely interesting to assess the effect of incorporating a bowhead-specific repeat library into the annotation process.

The alignments generated allow this to be examined in more detail. In the HOXA5 alignment (Figure 2.5), the human, mouse, cow and initial bowhead sequences exhibit a consistent start codon. However the published and final bowhead annotations in addition to the minke annotation contain a large amount of presumably spurious upstream sequence, amounting to approximately 400 and 1,000 bases in the cases of the bowhead annotations. A similar pattern can be observed in the alignments generated for HOXB2 (Figure 2.7) and HOXD12 (Figure 2.9). In each case, one or more of the bowhead annotations contain significant amounts of spurious upstream sequence with a premature start codon and occasionally additional exon sequence also. The situation with HOXD13 (Figure 2.10) appears somewhat more complex with apparent issues relating to all annotations with the notable exceptions of human and mouse. It is particularly significant that the cow sequence is

annotated as starting downstream of the consensus human and mouse start codon and has an invalid start codon. The minke and bowhead sequences also start significantly downstream of the human and mouse sequences and some contain an invalid start codon.

## Discussion

The genetic and molecular mechanisms by which longevity evolves remain largely unexplained. Given the declining costs of DNA sequencing, *de novo* genome sequencing is rapidly becoming affordable. The sequencing of genomes of long-lived species allows comparative genomics to be employed to study the evolution of longevity and has already provided candidate genes for further functional studies (de Magalhães and Keane, 2013). Nonetheless, deciphering the genetic basis of species differences in longevity has major intrinsic challenges (de Magalhães and Keane, 2013), and much work remains to uncover the underlying mechanisms by which some species live much longer than others. In this context, studying a species as long-lived and with such an extraordinary resistance to age-related diseases as the bowhead whale will help elucidate mechanisms and genes conferring longevity and disease resistance in mammals. Remarkably, large whales with over 1,000 times more cells than humans do not exhibit an increased cancer risk (Caulin and Maley, 2011), suggesting the existence of natural mechanisms that can suppress cancer more effectively in these animals. Having the genome sequence of the bowhead whale will allow researchers to study basic molecular processes and identify maintenance mechanisms that help preserve life, avoid entropy, and repair molecular damage. When compared to transcriptome data (Seim et al., 2014), the genome's greater completeness and quality permits additional (e.g. gene loss and duplication) and more thorough analyses. Besides, whereas the genomes of many commercially important agricultural species have been reported, the bowhead genome sequence is the first for a species key to a subsistence diet of indigenous communities. One of the outputs of this project will be to facilitate and drive research in this long-lived species. As such, the data are freely available to the scientific community on an online portal (<http://www.bowhead-whale.org/>), including the genome sequence, assembly, annotation and dN/dS results.

When considering the dN/dS results generated, it is necessary to be cognisant that some of the genes identified with dN/dS values exceeding 1 may not have experienced positive selection, as this could simply be due to chance fluctuations in dN and/or dS. Therefore both the statistical significance of these results and the substantial number of statistical tests employed would need to be accounted for before positive selection could be predicted with confidence. Similarly in the context of the analysis of unique bowhead residues, it is important to note that unique substitutions can be due simply to genetic drift and neutral changes, and are not necessarily indicative of positive selection. Furthermore, given that this analysis is preliminary, it would be necessary to verify any predictions using either data from either the transcriptome or from another closely-related whale species before they could be regarded with high confidence.

It is also important to take account of additional issues which can potentially confound the results of a genome sequencing project, including inaccurate annotation, genome sequence error and incompleteness. For example, an analysis of almost 3,000 orthologous protein-coding genes in seven terminal mammalian branches found that the inferred fraction of positively selected genes in sequences that were deficient in each of coverage, annotation and alignment was 7.2 times higher than that in genes with high trace sequencing coverage, “known” annotation status and perfect alignment scores (Schneider et al., 2009). The total sequence coverage for the bowhead genome is 154.3x with completeness of over 97%, which should result in minimal issues due to sequence errors and incompleteness.

However there did appear to be a large proportion of annotation errors in the published bowhead annotation, with ERCC1 (see Results) being a representative example. Furthermore, when the alignments of a number of additional genes which were initially identified in the analysis of positive selection were visually inspected, all contained suspicious novel bowhead sequence which proved spurious following manual annotation. This clearly indicates that without an accurate annotation, a genome-wide scan for selection degenerates into a scan for mis-annotation. In general, the annotation issues encountered primarily related to premature upstream start codons and intron-exon boundaries. It seems likely that these issues were at least in part the result of the settings used in MAKER2 and also the lack of training iterations using an *ab initio* gene predictor such as SNAP. In addition, it must be noted that a bowhead repeat library was not available which also could clearly have had a confounding effect on the results generated.

To explore this further, subsequent to publication I generated a new bowhead annotation. The initial run of MAKER2 used an altered configuration compared to the published version, with the settings to allow generation of annotations from the RNA-seq and protein data both enabled. I then used the protein lengths (Table 2.1) and alignments (Figures 2.5 – 2.10) of a set of 6 *Hox* genes to assess the extent to which this new annotation was superior to that published.

A very large number of genome-wide analyses of molecular evolution have now been reported in the literature, and the models typically employed to assess selective pressure are sensitive to errors on both foreground and background lineages. During the visual inspection of alignments generated for this project, suspicious gene annotations were observed for many species used in the analysis with the notable exceptions of human and mouse. As many species in addition to human and mouse are typically used in studies of molecular evolution, the accuracy and legitimacy of the results reported would therefore appear to be potentially questionable.

Furthermore, the fact that only coding sequences are usually assessed for evidence of selection raises a more fundamental question regarding the extent to which variation in highly complex traits such as longevity and cancer-resistance is due to selection on coding sequences. To illustrate, a recent study analysed correlations between gene functions and evidence for positive selection using a common statistical framework across several large surveys of coding and noncoding sequences throughout the human genome (Haygood et al., 2010). It was found that adaptation via coding changes was dominated by genes associated with immunity, olfaction, and male reproduction, whereas neutral development and function adapted mainly through noncoding changes. Indeed these associations have consistently dominated the large number of studies examining selection on coding sequences, which vastly exceed the number analysing non-coding sequences. In addition, it was observed that genes with highly tissue-specific expression underwent more adaptive coding changes, unlike adaptive noncoding changes, suggesting that pleiotropic constraints may inhibit such changes in broadly expressed genes. Regarding the evolution of longevity in particular, it appears implausible that it would be regulated by tissue-specific genes, which casts doubt on the relevance of analysing selection on coding sequences in this context. Given the difficulties associated with annotation of non-coding sequence however, it unfortunately does not appear feasible to pursue an analysis of selection at present, with the possible exception of proximal regions such as UTRs and promoters.

In conclusion, this project resulted in the genome sequence of the bowhead whale and the results of an annotation and comparative genome analysis being made available to the scientific community via an online portal. Although the genome sequence is high-coverage, the comparative analysis was confounded primarily by issues encountered during gene annotation. While the additional work on annotation which I undertook appears to have resulted in an improvement, however it is still clearly not at a level that would permit a robust and high-quality comparative analysis of selection to be completed. In terms of future work therefore, an obvious option would be to annotate the assembly again using a higher-quality pipeline. Ideally this would be accomplished using either the Ensembl (Flicek et al., 2013) or NCBI (<http://www.ncbi.nlm.nih.gov/books/NBK169439/>) annotation frameworks, although considerable patience may be required before this can be completed. It would also be important to include a bowhead-specific repeat library in this process. However as already stated, it is important to be aware that many gene annotations in various species which were generated by these pipelines were also observed to contain suspicious sequence during visual inspection of alignments. Indeed, human and mouse were the only species for which no suspicious annotations were noted. This is likely because these are the only two species which are currently included in the consensus coding sequence (CCDS) project (Pruitt et al., 2009). Therefore if a further

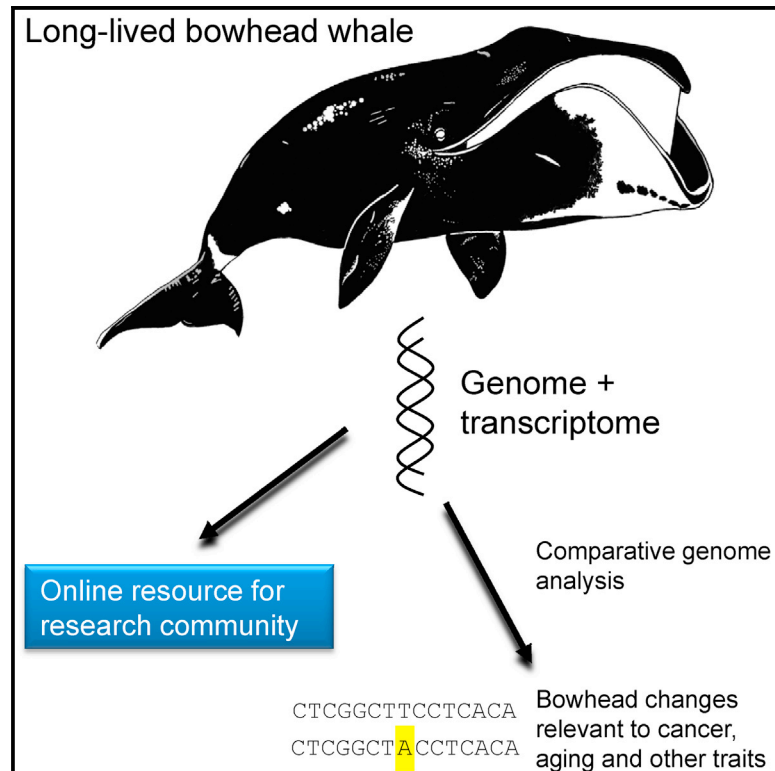
effort is undertaken to improve the bowhead (or indeed any) annotation for the purpose of subsequently completing a high-quality comparative genomics analysis, it should also be borne in mind that considerable work on the annotations of many of the other species used may likewise be necessary in order to generate high-confidence results. Given the amount of work that this would require, a collaborative effort by the entire community could possibly be the most realistic means of compiling a centralised database of manually validated gene annotations for as wide a range of species as possible.



# Cell Reports

## Insights into the Evolution of Longevity from the Bowhead Whale Genome

### Graphical Abstract



### Authors

Michael Keane, Jeremy Semeiks, ...,  
Bo Thomsen, João Pedro de Magalhães

### Correspondence

jp@senescence.info

### In Brief

The bowhead whale is the longest-lived mammal, possibly living over 200 years. Keane et al. sequence the bowhead genome and transcriptome and perform a comparative analysis with other cetaceans and mammals. Changes in bowhead genes related to cell cycle, DNA repair, cancer, and aging suggest alterations that may be biologically relevant.

### Highlights

- Genome and two transcriptomes of the bowhead whale, the longest-lived mammal
- Bowhead-specific mutations in genes associated with cancer and aging (e.g., ERCC1)
- Duplications in genes associated with DNA repair, cell cycle, and aging (e.g., PCNA)
- Changes in genes related to thermoregulation (UCP1) and other bowhead traits



Keane et al., 2015, *Cell Reports* 10, 112–122  
January 6, 2015 ©2015 The Authors  
<http://dx.doi.org/10.1016/j.celrep.2014.12.008>

CellPress

# Insights into the Evolution of Longevity from the Bowhead Whale Genome

Michael Keane,<sup>1,18</sup> Jeremy Semeiks,<sup>2,18</sup> Andrew E. Webb,<sup>3,18</sup> Yang I. Li,<sup>4,18,19</sup> Víctor Quesada,<sup>5,18</sup> Thomas Craig,<sup>1</sup> Lone Bruhn Madsen,<sup>6</sup> Sipko van Dam,<sup>1</sup> David Brawand,<sup>4</sup> Patrícia I. Marques,<sup>5</sup> Pawel Michalak,<sup>7</sup> Lin Kang,<sup>7</sup> Jong Bhak,<sup>8</sup> Hyung-Soon Yim,<sup>9</sup> Nick V. Grishin,<sup>2</sup> Nynne Hjort Nielsen,<sup>10</sup> Mads Peter Heide-Jørgensen,<sup>10</sup> Elias M. Oziolor,<sup>11</sup> Cole W. Matson,<sup>11</sup> George M. Church,<sup>12</sup> Gary W. Stuart,<sup>13</sup> John C. Patton,<sup>14</sup> J. Craig George,<sup>15</sup> Robert Suydam,<sup>15</sup> Knud Larsen,<sup>6</sup> Carlos López-Otín,<sup>5</sup> Mary J. O'Connell,<sup>3</sup> John W. Bickham,<sup>16,17</sup> Bo Thomsen,<sup>6</sup> and João Pedro de Magalhães<sup>1,\*</sup>

<sup>1</sup>Integrative Genomics of Ageing Group, Institute of Integrative Biology, University of Liverpool, Liverpool L69 7ZB, UK

<sup>2</sup>Howard Hughes Medical Institute and Departments of Biophysics and Biochemistry, University of Texas Southwestern Medical Center, Dallas, TX 75390-9050, USA

<sup>3</sup>Bioinformatics and Molecular Evolution Group, School of Biotechnology, Dublin City University, Glasnevin, Dublin 9, Ireland

<sup>4</sup>MRC Functional Genomics Unit, University of Oxford, Oxford OX1 3QX, UK

<sup>5</sup>Departamento de Bioquímica y Biología Molecular, Instituto Universitario de Oncología (IUOPA), Universidad de Oviedo, 33006 Oviedo, Spain

<sup>6</sup>Department of Molecular Biology and Genetics, Aarhus University, 8830 Tjele, Denmark

<sup>7</sup>Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA 24061, USA

<sup>8</sup>Personal Genomics Institute, Genome Research Foundation, Suwon 443-270, Republic of Korea

<sup>9</sup>KIOST, Korea Institute of Ocean Science and Technology, Ansan 426-744, Republic of Korea

<sup>10</sup>Greenland Institute of Natural Resources, 3900 Nuuk, Greenland

<sup>11</sup>Department of Environmental Science, Center for Reservoir and Aquatic Systems Research (CRASR) and Institute for Biomedical Studies, Baylor University, Waco, TX 76798, USA

<sup>12</sup>Department of Genetics, Harvard Medical School, Boston, MA 02115, USA

<sup>13</sup>The Center for Genomic Advocacy (TCGA) and Department of Biology, Indiana State University, Terre Haute, IN 47809, USA

<sup>14</sup>Department of Forestry and Natural Resources, Purdue University, West Lafayette, IN 47907, USA

<sup>15</sup>North Slope Borough, Department of Wildlife Management, Barrow, AK 99723, USA

<sup>16</sup>Battelle Memorial Institute, Houston, TX 77079, USA

<sup>17</sup>Department of Wildlife and Fisheries Sciences, Texas A&M University, College Station, TX 77843, USA

<sup>18</sup>Co-first author

<sup>19</sup>Present address: Department of Genetics, Stanford University, Stanford, CA 94305, USA

\*Correspondence: [jp@senescence.info](mailto:jp@senescence.info)

<http://dx.doi.org/10.1016/j.celrep.2014.12.008>

This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/3.0/>).

## SUMMARY

The bowhead whale (*Balaena mysticetus*) is estimated to live over 200 years and is possibly the longest-living mammal. These animals should possess protective molecular adaptations relevant to age-related diseases, particularly cancer. Here, we report the sequencing and comparative analysis of the bowhead whale genome and two transcriptomes from different populations. Our analysis identifies genes under positive selection and bowhead-specific mutations in genes linked to cancer and aging. In addition, we identify gene gain and loss involving genes associated with DNA repair, cell-cycle regulation, cancer, and aging. Our results expand our understanding of the evolution of mammalian longevity and suggest possible players involved in adaptive genetic changes conferring cancer resistance. We also found potentially relevant changes in genes related to additional processes,

including thermoregulation, sensory perception, dietary adaptations, and immune response. Our data are made available online (<http://www.bowhead-whale.org>) to facilitate research in this long-lived species.

## INTRODUCTION

The lifespan of some animals, including quahogs, tortoises, and certain whale species, is far greater than that of humans (Austad, 2010; Finch, 1990). It is remarkable that a warm-blooded species such as the bowhead whale (*Balaena mysticetus*) has not only been estimated to live over 200 years (estimated age of one specimen 211 SE 35 years), suggesting it is the longest-lived mammal, but also exhibits very low disease incidence until an advanced age compared to humans (George et al., 1999; Philo et al., 1993). As in humans, the evolution of longevity in whales was accompanied by low fecundity and longer developmental time (Tacutu et al., 2013), as predicted by evolutionary theory. The cellular, molecular, and genetic mechanisms underlying longevity and resistance to age-related diseases in bowhead

**Table 1. Statistics of the Bowhead Whale Genome Sequencing**

Sequence Data Generated			
Libraries	Total Data (Gb)	Sequence Coverage (for 2.91 Gb)	
200 bp paired-end	149.1	51.2×	
500 bp paired-end	141.7	48.7×	
3 kb mate-paired	57.3	19.7×	
5 kb mate-paired	72.5	24.9×	
10 kb mate-paired	28.5	9.8×	
<b>Total</b>	449.1	154.3×	
Genome Assembly Statistics			
Assembly	N50 (kb)	Number	Total Size (Gb)
Contigs	34.8	113,673	2.1
Scaffolds	877	7,227	2.3
See also <a href="#">Figures S1</a> and <a href="#">S2</a> .			

See also [Figures S1](#) and [S2](#).

whales are unknown, but it is clear that, in order to live so long, these animals must possess preventative mechanisms against cancer, immunosenescence, and neurodegenerative, cardiovascular, and metabolic diseases. In the context of cancer, whales, and bowhead whales, in particular, must possess effective antitumor mechanisms. Indeed, given their large size (in extreme cases adult bowhead whales can weigh up to 100 tons and are therefore among the largest whales) and exceptional longevity, bowhead whale cells must have a significantly lower probability of neoplastic transformation relative to humans ([Caulin and Maley, 2011](#); [de Magalhães, 2013](#)). Therefore, studying species such as bowhead whales that have greater natural longevity and resistance to age-related diseases than humans may lead to insights on the fundamental mechanisms of aging. Here, we report the sequencing and analysis of the genome of the bowhead whale, a species of the right whale family *Balaenidae* that lives in Arctic and sub-Arctic waters. This work provides clues regarding mechanisms underlying mammalian longevity and will be a valuable resource for researchers studying the evolution of longevity, disease resistance, and basic bowhead whale biology.

## RESULTS

### Sequencing and Annotation of the Bowhead Whale Genome

We sequenced the nuclear genome of a female bowhead whale (*Balaena mysticetus*) using the Illumina HiSeq platform at ~150× coverage. We followed established standards in the field in terms of sequencing paired-end libraries at high coverage plus mate-paired libraries of varying (3, 5, and 10 kb) insert sizes ([Table 1](#)). Contigs and scaffolds were assembled with ALLPATHS-LG ([Gnerre et al., 2011](#)). In line with other genomes sequenced with second-generation sequencing platforms, the contig N50 was 34.8 kb and scaffold N50 was 877 kb ([Table 1](#)); the longest scaffold in our assembly was 5,861 kb. In total, our assembly is ~2.3 Gb long. Genome size was estimated experimentally to be 2.91 Gb in another female and 2.87 Gb averaged with one male (see [Supplemental Results](#) and [Figure S1](#)), but this

discrepancy likely reflects highly repetitive regions, as observed for the genomes of other species with similar reported sizes such as the minke whale ([Yim et al., 2014](#)).

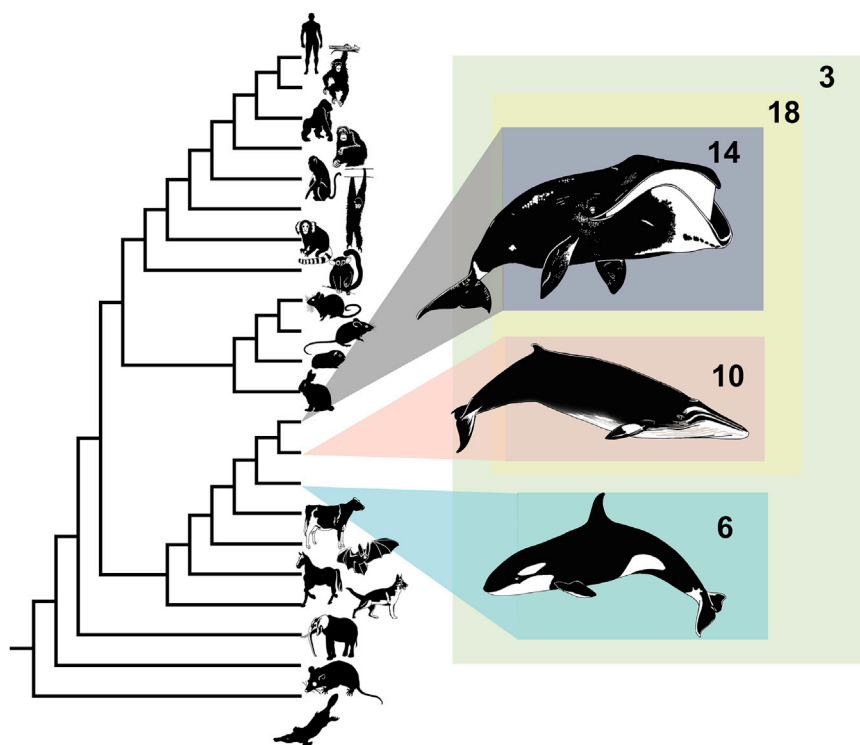
The full and partial completeness of the bowhead whale draft genome assembly was evaluated as 93.15% and 97.18%, respectively, by the CEGMA pipeline ([Parra et al., 2007](#)), which is comparable to the minke whale genome assembly ([Yim et al., 2014](#)). We also generated RNA sequencing (RNA-seq) data from seven adult bowhead whale tissues (cerebellum, kidney, muscle, heart, retina, liver, and testis) from specimens from Greenland and Alaska, resulting in two transcriptome assemblies (see [Experimental Procedures](#)) and annotated the genome using MAKER2, which combines ab initio methods, homology-based methods, and transcriptome data to derive gene models ([Holt and Yandell, 2011](#)). Our annotation contains 22,672 predicted protein-coding genes with an average length of 417 (median 307) amino acid residues. In addition, based on transcriptome data from two Alaskan individuals ([Table S1](#)), we estimated 0.5–0.6 SNPs per kilobase of RNA ([Table S2](#)). To begin annotation of the bowhead genome, we identified orthologs based on similarity with cow, human, and mouse genes/proteins (see [Experimental Procedures](#)), which allowed us to assign predicted gene symbols to 15,831 bowhead genes.

Moreover, to annotate microRNAs in the bowhead genome, we sequenced small RNA libraries prepared from kidney and skeletal muscle. The miRDeep algorithm ([Friedländer et al., 2008, 2012](#)) was used to integrate the sequencing data into a model of microRNA biogenesis by Dicer processing of predicted precursor hairpin structures in the genome, thus identifying 546 candidate microRNA genes. Of the 546 candidate miRNAs identified in the bowhead, 395 had seed sequences previously identified in miRNAs from human, cow, or mouse, whereas 151 did not. All of our data are available online from our Bowhead Whale Genome Resource portal (<http://www.bowhead-whale.org>).

### Analysis of the Draft Bowhead Whale Genome

Repeat sequences make up 41% of the bowhead genome assembly, most of which (78%) belong to the group of transposable elements (TEs). Although long interspersed nuclear elements (LINEs), such as L1, and short interspersed nuclear elements (SINEs) are widespread TEs in most mammalian lineages, the bowhead genome, similar to other cetacean genomes—minke, orca, and common bottlenose dolphin—is virtually devoid of SINEs ([Supplemental Folder 1](#)). LINE-1 (L1) is the most abundant TE, particularly in orca (90%) and minke whale (89%) ([Figure S2](#)). In comparison, TE diversity (measured with Shannon's index) in the bowhead genome (0.947) is higher than in orca (0.469) and minke whale (0.515) but lower than in dolphin (1.389) and cow ([Bovine Genome Sequencing and Analysis Consortium et al., 2009](#)) (1.534).

As a first assessment of coding genes that could be responsible for bowhead whale adaptations, we used bowhead coding sequences to calculate pairwise dN/dS ratios for 9,682, 12,685, and 11,158 orthologous coding sequences from minke whale (*Balaenoptera acutorostrata*), cow (*Bos taurus*), and dolphin (*Tursiops truncatus*), respectively. It is interesting to note that there are high levels of sequence conservation in the protein coding regions between bowhead and these species: 96% (minke),



**Figure 1. Phylogeny of Mammals Used in Codon-Based Maximum Likelihood Comparison of Selective Pressure Variation**

The number of candidate genes under positive selection on each lineage is indicated.

one copy in each species). We tested each of the extant whale lineages, the ancestral baleen whale, and the most recent common ancestor (MRCA) of bowhead, minke, and orca, a total of five lineages (Figure 1), for evidence of lineage-specific positive selection.

Of the two extant whales analyzed, the number of SGOs exhibiting signatures of lineage-specific positive selection were as follows: bowhead (15 gene families) and minke (ten gene families). The small number of candidates under positive selection likely reflects the high level of protein conservation between bowhead and other cetaceans as well as the stringent filtering of candidates due to data-quality concerns; all results and alignments are provided in Supplemental Folder 1. A few genes associated with disease were

92% (dolphin), and 91% (cow). This is not surprising, however, given the long generation time of cetaceans and of the bowhead whale, in particular, with animals only reaching sexual maturity at >20 years (Tacutu et al., 2013).

Because the minke whale is the closest relative to the bowhead (divergence time 25–30 million years ago [Gatesy et al., 2013]) with a sequenced genome and is smaller (<10 tons) and probably much shorter lived (maximum lifespan ~50 years) (Tacutu et al., 2013), comparisons between the bowhead and minke whale genomes may provide insights on the evolution of bowhead traits and of longevity, in particular. A number of aging- and cancer-associated genes were observed among the 420 predicted bowhead-minke orthologs with dN/dS exceeding 1, including *suppressor of cytokine signaling 2* (SOCS2), *apoptosis-inducing protein* (APT), *noggin* (NOG), and *leptin* (LEP). In addition, the top 5% genes with high dN/dS values for bowhead-minke relative to the values for minke-cow and minke-dolphin orthologs included *forkhead box O3* (FOXO3), *excision repair cross-complementing rodent repair deficiency, complementation group 3* (ERCC3), and *fibroblast growth factor receptor 1* (FGFR1). The data on dN/dS ratios are also available on our portal to allow researchers to do their own analysis and quickly retrieve gene(s) of interest.

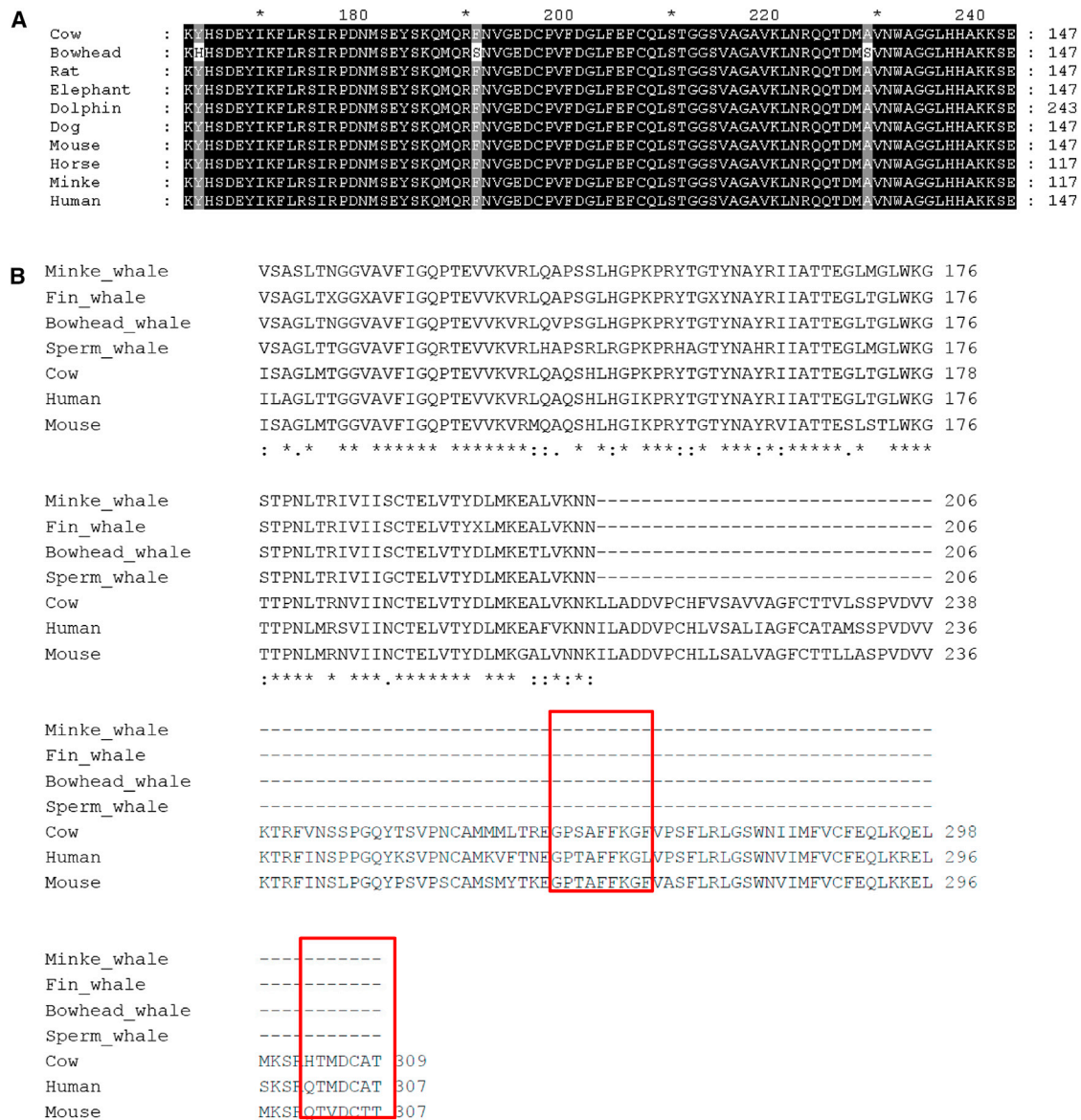
In a complementary and more detailed analysis of selective pressure variation, we used codon-based models of evolution (Yang, 2007) to identify candidate genes with evidence of lineage-specific positive selection (see Experimental Procedures). Using bowhead, minke, and orca protein-coding data along with a variety of available high-quality completed genomes from Laurasiatheria, Euarchontoglires, marsupial, and monotreme species, we identified a total of 866 single-gene ortholog families (SGOs) (i.e., these gene families have no more than

identified, including *BMP* and *activin membrane-bound inhibitor* (BAMBI), which has been associated with various pathologies, including cancer, and also poorly studied genes of potential interest like *GRB2-binding adaptor protein, transmembrane* (GAPT).

In addition to the codon-based models of evolution, we wished to identify bowhead whale specific amino acid replacement substitutions. To this end, we aligned orthologous sequences between the bowhead whale and nine other mammals—a total of 4,358 alignments (see Experimental Procedures). Lineage-specific residues identified in this way have previously been shown to be indicative of significant changes in protein function (Tian et al., 2013). Our analysis revealed several proteins associated with aging and cancer among the top 5% of unique bowhead residues by concentration (i.e., normalized by protein length), including ERCC1 (excision repair cross-complementing rodent repair deficiency, complementation group 1), HDAC1 (histone deacetylase 1), and HDAC2 (Figure 2A). ERCC1 is a member of the nucleotide excision repair pathway (Gillet and Schärer, 2006), and disruption results in greatly reduced lifespan in mice and accelerated aging (Weeda et al., 1997). Histone deacetylases play an important role in the regulation of chromatin structure and transcription (Lee et al., 1993) and have been associated with longevity in *Drosophila* (Rogina et al., 2002). As such, these represent candidates involved in adaptive genetic changes conferring disease resistance in the bowhead whale. The full results are available in Supplemental Folder 1.

In addition to genes related to longevity, several interesting candidate genes emerged from our analysis of lineage-specific residues of potential relevance to other bowhead traits. Of





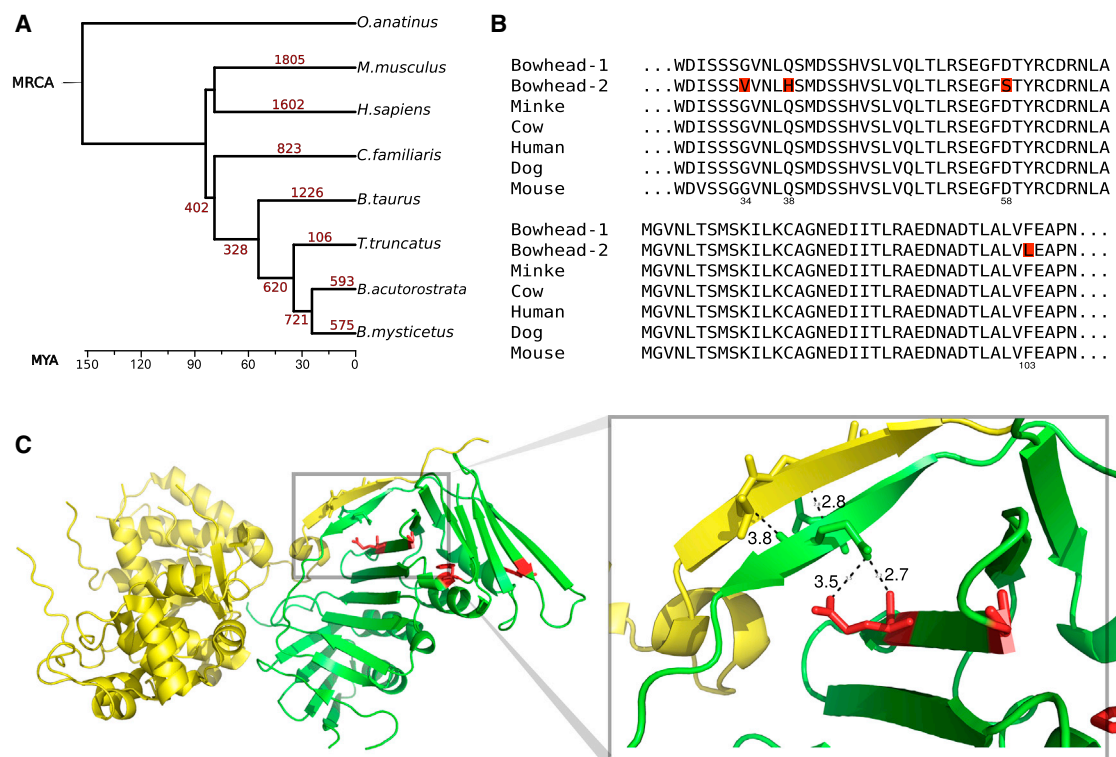
**Figure 2. Multiple Protein Sequence Alignments of HDAC2 and UCP1**

(A) Partial alignment of bowhead HDAC2 with mammalian orthologs. Unique bowhead residues are highlighted at human positions 68, 95, and 133. (B) Partial alignment of whale UCP1 with mammalian orthologs. Conserved regions involved in UCP1 are marked in red.

note, a number of proteins related to sensory perception of sound were also identified with bowhead-specific mutations, including otoraplin (OTOR) and cholinergic receptor, nicotinic, alpha 10 (CHRNA10), which could be relevant in the context of the bowhead's ability to produce high- and low-frequency tones simultaneously (Teruo et al., 2011). In addition, many proteins must play roles in the large differences in size and development between the bowhead and related species and our results reveal possible candidates for further functional studies; for example, in the top ten proteins, SNX3 (sorting nexin 3) has been associated in one patient with eye formation defects and microcephaly (Vervoort et al., 2002), and WDR5 (WD repeat-containing protein 5)

has been associated with osteoblast differentiation and bone development (Gori et al., 2006).

In the naked mole rat, a poikilotherm with a low metabolic rate and body temperature when compared to other mammals, unique changes in uncoupling protein 1 (UCP1), which is used to generate heat, have been previously found (Kim et al., 2011). Because the specific metabolic power output of cells in vivo for large whales must be much less than for smaller mammals (West et al., 2002), it is interesting to note that UCP1 of whales has a premature stop codon in C-terminal region, which is functionally important and conserved in other mammals (Figure 2B). It is tempting to speculate that these changes are related



**Figure 3. Gene Family Expansion and PCNA**

(A) Gene family expansion. Numbers in red correspond to the predicted number of gene expansion events during mammalian evolution. Mean divergence time estimates were used from TimeTree (Hedges et al., 2006) for scaling.

(B) Multiple sequence alignment of PCNA residues 28–107, showing bowhead whale-specific duplication (gene IDs: bmy 16007 and bmy 21945). Lineage-specific amino acids in the duplicated PCNA of bowhead whales are highlighted in red.

(C) Crystal structure of the PCNA (green) and FEN-1 (yellow) complex. Lineage-specific residues on the PCNA structure are colored in red. A zoom in on the structures reveals a putative interaction between two  $\beta$  sheets, one within PCNA and another within FEN-1. This interaction may be altered through a second interaction between the PCNA  $\beta$  sheet and a lineage-specific change from glutamine to histidine within PCNA. Distance measurements between pairs of atoms are marked in black. PDB accession number: 1UL1.

See also Table S3 and Figure S3.

to differences in thermoregulation between whales and smaller mammals.

### Potential Gene Duplications and Gene Losses

Gene duplication is a major mechanism through which phenotypic innovations can evolve (Holland et al., 1994; Kaessmann, 2010). Examples of mammalian phenotypic innovations associated to gene duplication include duplication of *RNASE1*, a pancreatic ribonuclease gene, in leaf-eating monkeys that contributed to adaptative changes in diet and digestive physiology (Zhang et al., 2002), a duplication of *GLUD1* in hominoids that subsequently acquired brain-specific functions (Burki and Kaessmann, 2004), and domestication of two syncytin gene copies that contributed to the emergence of placental development in mammals (Dupressoir et al., 2009). We surveyed the bowhead whale genome for expanded gene families that may reflect lineage-specific phenotypic adaptations and traits.

In the bowhead whale lineage, 575 gene families were predicted to have expanded (Figure 3). However, because gene expansion predictions are susceptible to false-positives owing to pseudogenes and annotation artifacts among other biases,

we applied a stringent filter based on percentage of identity (Experimental Procedures) that reduced the number of candidate expansions to 41 (see Supplemental Folder 1 for the complete list). A functional enrichment analysis of these gene families, using default parameters in DAVID (Huang et al., 2009), only revealed a statistically significant enrichment (after correction for multiple hypothesis testing; Bonferroni  $<0.001$ ) for genes associated with translation/ribosome. Given the association between translation and aging, for instance, in the context of loss of proteostasis (López-Otín et al., 2013), it is possible that these results reflect relevant adaptations in the bowhead whale.

Upon manual inspection of the gene expansion results, we found several duplicates of note. For instance, *proliferating cell nuclear antigen* (PCNA) is duplicated in bowhead whales with one copy harboring four lineage-specific residue changes (Figure 3B). Based on our RNA-seq data mapped to the genome (see Experimental Procedures and full results in Supplemental Folder 1), both PCNA copies are expressed in bowhead whale muscle, kidney, retina, and testis. By mapping the lineage-specific residues onto the structure of PCNA in complex with

FEN-1, we uncovered one amino acid substitution (Q38H), which may affect the interaction between PCNA and FEN-1 (Figure 3C). A subsequent branch-site test for selective pressure variation (see Experimental Procedures and Table S3) revealed that one substitution, D58S, may have undergone positive selection in the bowhead-whale lineage (with a posterior probability score of 0.983). The duplication of PCNA during bowhead-whale evolution is of particular interest due to its involvement in DNA damage repair (Hooge et al., 2002) and association with aging in that its levels in aged rat liver seem to relate to the decrease in the rate of cell proliferation (Tanno et al., 1996).

Another notable duplicated gene is *late endosomal/lysosomal adaptor, MAPK and MTOR activator 1* (LAMTOR1), in which six bowhead-specific amino acid changes were identified (Figure S3). LAMTOR1 is involved in amino acid sensing and activation of mTORC1, a gene strongly associated with aging and cancer (Cornu et al., 2013). The original LAMTOR1 copy was expressed in all bowhead whale adult tissues for which we have data, with the duplicate having much lower (but detectable) expression in heart and retina. Also of note, putative duplications of *26S proteasome non-ATPase regulatory subunit 4* (PSMD4) and *ubiquitin carboxyl-terminal esterase L3* (UCHL3) were identified with evidence of expression, which is intriguing considering the known involvement of the proteasome-ubiquitin system in aging (López-Otín et al., 2013) and given previous evidence that this system is under selection specific to lineages where longevity increased (Li and de Magalhães, 2013); UCHL3 has also been involved in neurodegeneration (Kurihara et al., 2001). Other gene duplications of potential interest for their role in mitosis, cancer, and stress response include *cAMP-regulated phosphoprotein 19* (ARPP19), which has three copies even though we only detected expression of two copies, *stomatolike 2* (STOML2), *heat shock factor binding protein 1* (HSBP1) with four copies of which two appear to be expressed, *spermine synthase* (SMS) and *suppression of tumorigenicity 13* (ST13).

Similar to previous genome characterizations, we chose the complete set of known protease genes for a detailed supervised analysis of gene loss (Quesada et al., 2009). This procedure highlighted multiple gene loss events potentially related to the evolution of several cetacean traits, including adaptations affecting the immune system, blood homeostasis, digestive system, and dentition (Figure S4). Thus, the cysteine protease CASP12, a modulator of the activity of inflammatory caspases, has at least one conserved premature stop codon in bowhead and minke whales. Interestingly, whereas this protease is conserved and functional in almost all of the terrestrial mammals, most human populations display different deleterious variants (Fischer et al., 2002), presumably with the same functional consequences as the premature stop codons in whales. Likewise, two paralogues of carboxypeptidase A (CPA2 and CPA3) have been pseudogenized in bowhead and minke whales. Notably, CPA variants have been associated with increased risk for prostate cancer in humans (Ross et al., 2009), which could be of interest in the context of reduced cancer susceptibility in whales compared with humans (de Magalhães, 2013).

Additionally, we found that multiple coagulation factors have been lost in bowhead and minke whales. The finding of bowhead whale-specific changes is also noteworthy because it could be

related to the special characteristics of this mammal. For example, OTUD6A, a cysteine protease with a putative role in the innate immune system (Kayagaki et al., 2007), is specifically lacking in the assembled genome and expressed sequences of the bowhead whale. In addition, whereas the enamel metalloprotease MMP20 has been lost in bowhead and minke whales (Yim et al., 2014), our analysis suggests that these genomic events happened independently (see alignments in Supplemental Folder 1). Finally, as aforementioned, the cysteine protease UCHL3 seems to have been duplicated through a retrotranscription-mediated event in a common ancestor to bowhead and minke whales, although only the genome of the bowhead whale shows a complete, putatively functional open reading frame for this extra copy of the gene. UCHL3 may play a role in adipogenesis (van Beekum et al., 2012), which indicates that this duplication might be related to the adaptation of the bowhead whale to the challenging arctic environment. These results suggest specific scenarios for the role of proteolysis in the evolution of *Mysticetes*. Specifically, given the relationship between immunity and aging (López-Otín et al., 2013), some of these findings might open new approaches for the study of this outstanding cetacean.

## DISCUSSION

The genetic and molecular mechanisms by which longevity evolves remain largely unexplained. Given the declining costs of DNA sequencing, de novo genome sequencing is rapidly becoming affordable. The sequencing of genomes of long-lived species allows comparative genomics to be employed to study the evolution of longevity and has already provided candidate genes for further functional studies (de Magalhães and Keane, 2013). Nonetheless, deciphering the genetic basis of species differences in longevity has major intrinsic challenges (de Magalhães and Keane, 2013), and much work remains to uncover the underlying mechanisms by which some species live much longer than others. In this context, studying a species so long lived and with such an extraordinary resistance to age-related diseases as the bowhead whale will help elucidate mechanisms and genes conferring longevity and disease resistance in mammals. Remarkably, large whales with over 1,000 times more cells than humans do not exhibit an increased cancer risk (Caulin and Maley, 2011), suggesting the existence of natural mechanisms that can suppress cancer more effectively in these animals. Having the genome sequence of the bowhead whale will allow researchers to study basic molecular processes and identify maintenance mechanisms that help preserve life, avoid entropy, and repair molecular damage. When compared to transcriptome data (Seim et al., 2014), the genome's greater completeness and quality permits additional (e.g., gene loss and duplication) and more thorough analyses. Besides, whereas the genomes of many commercially important agricultural species have been reported, the bowhead genome sequence is the first for a species key to a subsistence diet of indigenous communities. One of the outputs of this project will be to facilitate and drive research in this long-lived species. Data and results from this project are thus made freely available to the scientific community on an online portal (<http://www.bowhead-whale.org/>). We provide

this key resource for studying the bowhead whale and its various traits, including its exceptional longevity and resistance to diseases.

## EXPERIMENTAL PROCEDURES

### DNA and RNA Sampling in Greenland

Bowhead (*Balaena mysticetus*) DNA used for genome sequencing was isolated from muscle tissue sampled from a 51-year-old female (ID no. 325) caught in the Disko Bay, West Greenland in 2009 (Heide-Jørgensen et al., 2012). Tissue samples were stored at  $-20^{\circ}\text{C}$  immediately after collection. Age estimation was performed using the aspartic acid racemization technique (Garde et al., 2007). CITES no. 12GL1003387 was used for transfer of biological material. Bowhead RNA used for RNA-seq and small RNA analysis was isolated from two different individuals: kidney samples were from a 44-year-old female (ID no. 500) and muscle samples were isolated from a 44-year-old male (ID no. 322). For more details of the individual whales, see Heide-Jørgensen et al. (2012).

### Genome Sequencing

DNA was extracted following standard protocols, quantified using Qubit and run on an agarose gel to ensure no degradation had occurred. We then generated  $\sim 150\times$  coverage of the genome using the Illumina HiSeq 2000 platform with 100 bp reads, sequencing paired-end libraries, and mate-paired libraries with insert sizes of 3, 5, and 10 kb (Table 1). Sequencing was performed at the Liverpool Centre for Genomic Research (CGR; <http://www.liv.ac.uk/genomic-research/>).

### Genome Assembly

Libraries were preprocessed in-house by the CGR to remove adaptor sequences. The raw fastq files were trimmed for the presence of the Illumina adaptor sequence using Cutadapt and then subjected to window-based quality trimming using Sickle with a minimum window quality score of 20. A minimum read-length filter of 10 bp was also applied. Libraries were then assembled with ALLPATHS-LG (Gnerre et al., 2011), which performed all assembly steps including read error correction, initial read alignment, and scaffolding. ALLPATHS-LG build 43762 was used with the default input parameters, including  $K = 96$ . Several build parameters were automatically determined by the software at run time per its standard algorithm. Of  $2.88 \times 10^9$  paired fragment reads and  $1.87 \times 10^9$  paired jumping reads, 0.015% were removed as poly(A) and 1.5% were removed due to low-frequency kmers; 54% of jumping read pairs were error-corrected, and overall 33% of jumping pairs were redundant. In total, we used 216 Gbp for the 2.3 Gb assembly, meaning that coverage retained for the assembly was  $\sim 95\times$ . Full assembly and read usage data are shown in Supplemental Folder 2. Assembly completeness was assayed with CEGMA by searching for 248 core eukaryotic genes (Parra et al., 2007).

### Genome Size Determination

To determine the genome size for bowhead whale, spleen tissues were acquired from one male (10B17) and one female (10B18). Both whales were harvested in 2010 as part of the native subsistence hunt in Barrow, Alaska. Sample processing and staining followed the methods of Vindeløv and Christensen (1994). Instrument description and additional methodological details are provided in Ozolator et al. (2014). Briefly, flow cytometric genome size determination is based on propidium iodide fluorescent staining of nuclear DNA. Mean fluorescence is calculated for cells in the G0 and G1 phases of the cell cycle. This method requires direct comparison to known standards to convert measured fluorescence to pg of DNA. The primary standard used in this study was the domestic chicken (*Gallus gallus domesticus*). Chicken red blood cells are widely used as a genome size standard, with an accepted genome size of  $C = 1.25$  pg. Chicken whole blood was purchased from Innovative Research. Mouse (*Mus musculus*) and rat (*Rattus norvegicus*) were included as internal checks, with estimates for both falling within 3% of previously published genome size estimates (Vinogradov, 1998). Spleen tissues from three male 129/SvEvTac laboratory mice and a single male Harlan SD Sprague-Dawley laboratory rat were used.

### Transcriptome Sequencing and Assembly: Greenland Samples

Total RNA was extracted from the kidney and muscle employing the mirVanaTM RNA extraction kit (Ambion). RNA integrity of the individual RNA samples was assessed on a 1% agarose gel using an Agilent 2100 Bioanalyzer (Agilent Technologies). Library preparation was performed using the ScriptSeqTM mRNA-seq library preparation kit from Epicenter according to the manufacturer's protocol (Epicenter) and sequenced (100 bp paired end) as multiplexed samples using the Illumina HiSeq 2000 analyzer. Fastq generation and demultiplexing were performed using the CASAVA 1.8.2 package (Illumina). The fastq files were filtered for adapters, quality, and length using Trimmomatic (v.0.27), with a window size of 4, a base quality cutoff of 20, and a minimum length of 60 (Lohse et al., 2012). De novo transcriptome assembly was performed using the short read assembler software Trinity (release 2013-02-25), which is based on the de Bruijn graph method for assembly, with default settings (Grabherr et al., 2011).

### Transcriptome Sequencing and Assembly: Alaskan Samples

Tissue biopsies were obtained from two male bowhead whales harvested by Inupiat hunters at Barrow, Alaska during the Fall hunt of 2010; heart, cerebellum, liver, and testes were biopsied from male bowhead number 10B16, and retina from male bowhead 10B20. Samples were immediately placed in liquid nitrogen and transported in a dry shipper to Purdue University. RNA was extracted using TRIzol reagent (Invitrogen) following the manufacturer's protocol. RNA was purified using an Invitrogen PureLink Micro-to-Midi columns from the Total RNA Purification System using the standard protocol. RNA quantity and quality was estimated with a spectrophotometer (Nanodrop) and by gel electrophoresis using an Agilent model 2100 Bioanalyzer. cDNA libraries were constructed by random priming of chemically sheared poly A captured RNA. Randomly primed DNA products were blunt ended. Products from 450–650 bp were then isolated using a PippenPrep. After the addition of an adenine to the fragments, a Y primer amplification was used to produce properly tailed products. Paired-end sequences of 100 bp per end were generated using the Illumina HiScan platform. Sequences with primer concatamers, weak signal, and/or poly A/T tails were culled. The Trinity software package for de novo assembly (Grabherr et al., 2011) was used for transcript reconstruction (Table S1).

### Small RNA Sequencing and Annotation

To annotate microRNA genes in the bowhead genome, we conducted deep sequencing of two small RNA libraries prepared from muscle and kidney tissues (Greenland samples). Total RNA was isolated using mirVana miRNA Isolation Kit (Ambion). Small RNA in the 15–40 nucleotides range was gel purified and small RNA libraries were prepared for next-generation sequencing using the ScriptMiner Small RNA-Seq Library Preparation Kit (Epicenter). The two libraries were sequenced on an Illumina Hi-Seq 2000 instrument to generate single end sequences of 50 nucleotides. Primary data analysis was done using the Illumina CASAVA Pipeline software v.1.8.2, and the sequence reads were further processed by trimming for adapters and filtering for low quality using Trimmomatic (Lohse et al., 2012). Identification of conserved and novel candidate microRNA genes in the bowhead genome was accomplished by applying the miRDeep2 algorithm (Friedländer et al., 2008, 2012).

### Evaluation of Repeat Elements

To evaluate the percentage of repeat elements, RepeatMasker (v.4.0.3; <http://www.repeatmasker.org/>) was used to identify repeat elements, with parameters set as “-s -species mammal.” RMBlast was used as a sequence search engine to list out all types of repeats. Percentage of repeat elements was calculated as the total number of repeat region divided by the total length of the genome, excluding the N-region. Genomes of minke whale (*Balaenoptera acutorostrata*), orca (*Orcinus orca*), common bottlenose dolphin (*Tursiops truncatus*), and cow (*Bos taurus*) were downloaded from NCBI and run in parallel for comparison with the bowhead genome.

### Genome Annotation

Putative genes were located in the assembly by structural annotation with MAKER2 (Holt and Yandell, 2011), which combined both bowhead



transcriptomes with comparative and de novo prediction methods including BLASTX, Exonerate, SNAP, Genemark, and Augustus. In addition to the RNA-seq data, the entire SwissProt database and the draft proteome of dolphin were used as input to the comparative methods. Repetitive elements were found with RepeatMasker (<http://www.repeatmasker.org/>). The complete set of MAKER input parameters, including training sets used for the de novo prediction methods, are listed in [Supplemental Folder 2](#). In total, 22,672 protein-coding genes were predicted with an average length of 417 (median 307) amino acid residues.

The RNA-seq data from seven adult bowhead tissues described above were then mapped to the genome: FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) was used for quality control to make sure that data of all seven samples was of acceptable quality. STAR (Dobin et al., 2013) was used to generate genome files from the bowhead assembly and to map the reads to the bowhead genome with 70.3% of reads mapping, which is in line with other results including those in the minke whale (Yim et al., 2014). To count the reads overlapping genes, we used ReadCounter (van Dam et al., 2015). The results obtained from all seven samples were combined into a single file describing the number of nonambiguously mapping reads for each gene (full results in [Supplemental Folder 1](#)). Of the 22,672 predicted protein-coding genes, 89.5% had at least ten reads mapping and 97.5% of predicted genes had at least one read mapping to them, which is again comparable to other genomes like the minke whale genome (Yim et al., 2014).

To allow the identification of orthologous relationships with bowhead proteins, all cow protein sequences were downloaded from Ensembl (Flicek et al., 2013). Cow was initially used because it is the closest relative to the bowhead with a high-quality annotated genome available. First, BLASTP ( $10^{-5}$ ) was used to find the best hit in the cow proteome for every predicted bowhead protein, and then the reciprocal best hit for each cow protein was defined as an ortholog. In addition, human and mouse orthologs from the OPTIC pipeline (see below) were used to assign predicted gene symbols to genes and proteins. A total of 15,831 bowhead genes have a putative gene symbol based on these predictions. Homologs in minke whale and dolphin were also derived and are available on our bowhead genome portal.

### Genome Portal

To facilitate further studies of these animals, we constructed an online genome portal: The Bowhead Whale Genome Resource (<http://www.bowhead-whale.org/>). Its database structure, interface, and functionality were adapted from our existing Naked Mole Rat Genome Resource (Keane et al., 2014). Our data and results are available from the portal, and supplemental methods and data files are also available on GitHub (<https://github.com/maglab/bowhead-whale-supplementary>).

### Pairwise dN/dS Analysis

The CodeML program from the PAML package was used to calculate pairwise dN/dS ratios (Yang, 2007). This is done using the ratio of nonsynonymous substitutions per nonsynonymous site (dN) to synonymous substitutions per synonymous site (dS), dN/dS, or  $\omega$  (Yang, 2007). Specifically, these pairwise dN/dS ratios were calculated for bowhead coding sequences and orthologous sequences from minke, cow, and dolphin, excluding coding sequences that were less than 50% of the length of the orthologous sequence. The results were then ranked by decreasing dN/dS and are available on our bowhead genome portal. In addition, the ratio of the bowhead-minke dN/dS value to the higher of the dN/dS values for minke-cow and minke-dolphin was calculated to identify genes that evolved more rapidly on the bowhead lineage.

### Assessment of Selective Pressure Variation across Single-Genes Orthologous Families Using Codon-Based Models of Evolution

To accurately assess variation in selective pressure on the bowhead, minke, and orca lineages in comparison to extant terrestrial mammals, we created a protein-coding database spanning the placental mammals. Along with the orca (<http://www.ncbi.nlm.nih.gov/bioproject/189949>), minke (Yim et al., 2014), and bowhead data described above, we extracted protein coding sequences from Ensembl Biomart v.73 (Flicek et al., 2013) for the following

18 genomes: chimpanzee, cow, dog, elephant, gibbon (5.6 $\times$  coverage), gorilla, guinea pig, horse, human, macaque, marmoset, microbat, mouse, opossum, orangutan, platypus, rabbit, and rat. These genomes were all high coverage (mostly >6 $\times$  coverage) with the exception of gibbon ([Supplemental Folder 2](#)). Sequence similarity searches were performed using mpi-BLAST (v 1.6.0) (Altschul et al., 1990) (<http://www.mpiblast.org/>) on all proteins using a threshold of  $10^{-7}$ . Gene families were identified using in-house software that clusters genes based on reciprocal BLAST hits (Altschul et al., 1990). We identified a total of 6,630 gene families from which we extracted the single-gene orthologous families (SGOs). Families were considered SGOs if we identified a single-gene representative in each species (one-to-one orthologs), and to account for lower coverage genomes and missing data we also considered cases where a specific gene was not present in a species, i.e., one-to-zero orthology. SGOs were only considered for subsequent analysis if they contained more than seven species in total and if they contained no internal stop codons (indicative of sequencing errors). In total, we retained 866 SGOs for further analysis. Multiple sequence alignments (MSAs) were generated using default parameters in PRANK (v.100802) (Löytynoja and Goldman, 2008). To minimize potential false-positives due to poor sequence quality, the MSAs of the 866 SGOs underwent strict data-quality filtering. The first filter prohibited the presence of gaps in the MSA if created by unique insertions (>12 bp) in either bowhead or minke sequences. The second filter required unaligned bowhead or minke sequences to be at least half the length of their respective MSA. These two filters refined the number of testable SGOs to 319. The gene phylogeny of each SGO was inferred from the species phylogeny (Morgan et al., 2013). CodeML from the PAML software package (v.4.4e) (Yang, 2007) was employed for our selective pressure variation analyses. We analyzed each of the 319 refined SGOs using the nested codon-based models of evolution under a maximum likelihood framework. We employed the likelihood ratio test (LRT) using nested models of sequence evolution to evaluate a variety of models of codon sequence evolution (Yang, 2007). In general, these codon models allow for variable dN/dS ratios (referred to as  $\omega$  throughout) among sites in the alignment, along different lineages on our phylogenetic tree, or a combination of both variations across lineages and sites. To assess the significance of fit of each model to the data, we used the recommended LRTs in CodeML (Yang, 2007) for comparing nested models (see [Supplemental Folder 2](#)). The LRT test statistic approximates the chi-square ( $\chi^2$ ) distribution critical value with degrees of freedom equal to the number of additional free parameters in the alternative model. The goal of the codon-based modeling is to determine the selective pressures at work in a lineage and site-specific manner.

The models applied follow the standard nomenclature (i.e., model M1, M2, A, and A null) (Yang, 2007). Model M1 assumes that there are two classes of sites—those with an  $\omega$  value of zero and those with an  $\omega$  value of 1. Model M2 allows for three classes of sites—one with an  $\omega$  value of zero, one with an  $\omega$  value of one and one with an  $\omega$  value that is not fixed to any value. Given the relationship between M1 and M2, they can be tested for the significance of the difference of the fit of these two models using an LRT with df = 2. Finally, we used model A that allows the  $\omega$  value to vary across sites and across different lineages in combination. With model A, we can estimate the proportion of sites and the dN/dS ratio in the foreground lineage of interest in comparison to the background lineages and the estimated dN/dS ratio is free to vary above 1 (i.e., positive selection). Model A can be compared with its site-specific counterpart (model M1) using the LRT with df = 2. In addition, the lineage and site-specific model model A null was applied as a second LRT with model A. In model A null, the additional site category is fixed at neutral rather than being estimated from the data, and this LRT provides an additional test for model A (Zhang et al., 2005). In this way, we performed independent tests on each of the extant cetacean lineages (orca, minke, and bowhead), as well as testing each ancestral cetacean branch (the MRCA of the two baleen whales and the MRCA of all three cetaceans), to determine if there were signatures of positive selection that are unique to each lineage (Yang and dos Reis, 2011). Using empirical Bayesian estimations, we identified the specific residues that are positively selected in each lineage tested. Positive selection was inferred if all of the following criteria were met: (1) if the LRT was significant, (2) if the parameters estimated under that model were concurrent with positive selection, and (3) if the alignment in that region was of high quality (as judged by alignment

completeness and quality in that region). The posterior probability (PP) of a positively selected site is estimated using two calculations: Naive Empirical Bayes (NEB) or Bayes Empirical Bayes (BEB) (Yang, 2007). If both NEB and BEB are predicted, we reported the BEB results as they have been shown to be more robust under certain conditions (Yang et al., 2005). For all models used in the analysis where  $\omega$  is estimated from the data, a variety of starting  $\omega$  values was used for the calculation of likelihood estimates. This ensures that the global minimum is reached.

### Identification of Proteins with Bowhead-Unique Residues

An in-house Perl pipeline was used to align each bowhead protein with orthologs from nine other mammals: human (*Homo sapiens*), dog (*Canis familiaris*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), minke whale (*Balaenoptera acutorostrata*), cow (*Bos taurus*), dolphin (*Tursiops truncatus*), horse (*Equus caballus*), and elephant (*Loxodonta africana*) and then identify the unique bowhead amino acid residues. Gaps were excluded from the analysis, and a maximum of one unknown residue was allowed in species other than the bowhead. The results were ranked by the number of unique residues normalized by the protein length (full results in Supplemental Folder 1).

### Gene Expansion Analysis, Filtering, and Expression

Human, mouse, dog, cow, dolphin, and platypus genomes and gene annotations were obtained from Ensembl (Flicek et al., 2013), the genome and gene annotation of minke whale were obtained from Yim et al. (2014). In total, 21,069, 22,275, 19,292, 19,988, 15,769, 17,936, 20,496, and 22,733 human, mouse, dog, cow, dolphin, platypus, minke whale, and bowhead whale genes, respectively, were used to construct orthology mappings using OPTIC (Heger and Ponting, 2007). Briefly, OPTIC builds phylogenetic trees for gene families by first assigning orthology relationships based on pairwise orthologs computed using PhyOP (Goodstadt and Ponting, 2006). Then, a tree-based method, PhyOP, is used to cluster genes into orthologous groups, and, last, gene members are aligned and phylogenetic trees built with TreeBeST (Vilella et al., 2009). Further details are available in the OPTIC paper (Heger and Ponting, 2007). Predicted orthology groups can be accessed at [http://genserv.anat.ox.ac.uk/clades/vertebrates\\_bowhead](http://genserv.anat.ox.ac.uk/clades/vertebrates_bowhead).

To identify gene families that underwent expansion, gene trees were reconciled with the consensus species tree, and duplicated nodes were identified. The tree used, derived from TimeTree (Hedges et al., 2006), was: (mm\_oanatinus5, ((mm\_cfamiliaris3, (mm\_btaurus, (mm\_ttruncatus, (mm\_balaenoptera, mm\_bmysticetus))), (mm\_hsapiens10, mm\_mmusculus5))). The following algorithm was used to reconcile gene and species trees.

A stringent filter was applied to the data so that gene duplicates in bowhead whales were required to differ by at most 10% in protein sequence from a cognate copy but were also required to differ by at least 1% to avoid assembly artifacts and to remove recently duplicated copies with no function. Further manual inspection of the alignments was performed. Gene expression inferred from our RNA-seq data was used to check the expression of duplicates.

An in-house peptide-sensitive approach was used to align the PCNA cDNA into codons, and CodeML/PAML was used to test M0, a one-rate model that assumes the same rate of evolution in all branches against M2<sup>a</sup>, a branch site test with one rate for the background and one rate for the bowhead whale branch (Yang, 2007).

### ACCESSION NUMBERS

Our data and results can be downloaded from the Bowhead Whale Genome Resource (<http://www.bowhead-whale.org/downloads/>). In addition, data are available at the NCBI BioProject PRJNA194091 with raw sequencing reads in the Sequence Read Archive (SRP050351).

### SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Results, four figures, three tables, and two supplemental data files and can be found with this article online at <http://dx.doi.org/10.1016/j.celrep.2014.12.008>.

### AUTHOR CONTRIBUTIONS

G.M.C., J.C.G., R.S., J.W.B., B.T., and J.P.M. conceived and designed the study; L.B.M., E.M.O., and C.W.M. performed the experiments; M.K., J.S., A.E.W., Y.I.L., V.Q., L.B.M., S.v.D., D.B., P.I.M., P.M., L.K., J.B., H.-S.Y., G.W.S., J.C.P., C.L.-O., M.J.O., J.W.C., and J.P.M. analyzed the data; T.C., N.V.G., N.H.N., M.P.H.-J., R.S., and K.L. contributed reagents/materials/analysis tools; and M.K. and J.P.M. wrote the paper.

### ACKNOWLEDGMENTS

This project was supported by grants from the Life Extension Foundation and the Methuselah Foundation to J.P.M. We thank the Inuit whaling captains of the Alaska Eskimo Whaling Commission and the Barrow Whaling Captains' Association for allowing us to sample their whales and their willingness to support the transcriptome study. This study was also partly funded by the Augustinus Foundation to K.L. T.C. was supported by a Wellcome Trust grant (WT094386MA) to J.P.M. and M.K. was supported by a studentship from the University of Liverpool's Faculty of Health and Life Sciences. J.S. was supported by grants from the NIH (GM094575 to N.V.G.) and the Welch Foundation (I-1505 to N.V.G.). Y.I.L. was supported by a Nuffield Department of Medicine Prize studentship from the University of Oxford. C.L.-O. is an Investigator of the Botin Foundation also supported by grants from Ministerio de Economía y Competitividad-Spain and Instituto de Salud Carlos III (RTICC)-Spain. M.J.O. and A.E.W. are funded by Science Foundation Ireland Research Frontiers Programme Grant (EOB2763) to M.J.O. M.J.O. would also like to acknowledge the Fulbright Commission for the Fulbright Scholar Award 2012-2013. M.J.O. and A.E.W. thank the SFI/HEA Irish Centre for High-End Computing (ICHEC) for processor time and technical support. This work was also supported by the Korea Institute of Ocean Science and Technology (KIOST) in-house program (PE99212). Further thanks to Prof. Chris Ponting and Dr. Andreas Heger for hosting gene orthology predictions from OPTIC, the University of Liverpool High Performance Computing facilities for processor time, Eric de Sousa for help with RNA-seq data analysis, and to Louise Crompton for assistance in compiling and formatting the bibliography. Last, we are grateful to the staff at the Liverpool Centre for Genomic Research for advice during this project.

Received: September 7, 2014

Revised: November 21, 2014

Accepted: December 3, 2014

Published: December 24, 2014

### REFERENCES

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.
- Austad, S.N. (2010). Methuselah's Zoo: how nature provides us with clues for extending human health span. *J. Comp. Pathol.* 142 (Suppl 1), S10–S21.
- Burki, F., and Kaessmann, H. (2004). Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. *Nat. Genet.* 36, 1061–1063.
- Caulin, A.F., and Maley, C.C. (2011). Peto's Paradox: evolution's prescription for cancer prevention. *Trends Ecol. Evol.* 26, 175–182.
- Bovine Genome Sequencing and Analysis Consortium, Elsik, C.G., Tellam, R.L., Worley, K.C., Gibbs, R.A., Muzny, D.M., Weinstock, G.M., Adelson, D.L., Eichler, E.E., Elnitski, L., et al. (2009). The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* 324, 522–528.
- Cornu, M., Albert, V., and Hall, M.N. (2013). mTOR in aging, metabolism, and cancer. *Curr. Opin. Genet. Dev.* 23, 53–62.
- de Magalhães, J.P. (2013). How ageing processes influence cancer. *Nat. Rev. Cancer* 13, 357–365.
- de Magalhães, J.P., and Keane, M. (2013). Endless paces of degeneration—applying comparative genomics to study evolution's moulding of longevity. *EMBO Rep.* 14, 661–662.

- Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21.
- Dupressoir, A., Vernochet, C., Bawa, O., Harper, F., Pierron, G., Opolon, P., and Heidmann, T. (2009). Syncytin-A knockout mice demonstrate the critical role in placentalization of a fusogenic, endogenous retrovirus-derived, envelope gene. *Proc. Natl. Acad. Sci. USA* 106, 12127–12132.
- Finch, C. (1990). *Longevity, Senescence, and the Genome* (Chicago: University of Chicago Press).
- Fischer, H., Koenig, U., Eckhart, L., and Tschachler, E. (2002). Human caspase 12 has acquired deleterious mutations. *Biochem. Biophys. Res. Commun.* 293, 722–726.
- Flicek, P., Ahmed, I., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., et al. (2013). Ensembl 2013. *Nucleic Acids Res.* 41, D48–D55.
- Friedländer, M.R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., and Rajewsky, N. (2008). Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.* 26, 407–415.
- Friedländer, M.R., Mackowiak, S.D., Li, N., Chen, W., and Rajewsky, N. (2012). miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.* 40, 37–52.
- Garde, E., Heide-Jørgensen, M.P., Hansen, S.H., Nachman, G., and Forchhammer, M.C. (2007). Age-specific growth and remarkable longevity in narwhals (*Monodon monoceros*) from West Greenland as estimated by aspartic acid racemization. *J. Mammal.* 88, 49–58.
- Gatesy, J., Geisler, J.H., Chang, J., Buell, C., Berta, A., Meredith, R.W., Springer, M.S., and McGowen, M.R. (2013). A phylogenetic blueprint for a modern whale. *Mol. Phylogenet. Evol.* 66, 479–506.
- George, J.C., Bada, J., Zeh, J., Scott, L., Brown, S.E., O'Hara, T., and Suydam, R. (1999). Age and growth estimates of bowhead whales (*Balaena mysticetus*) via aspartic acid racemization. *Can. J. Zool.* 77, 571–580.
- Gillet, L.C.J., and Schärer, O.D. (2006). Molecular mechanisms of mammalian global genome nucleotide excision repair. *Chem. Rev.* 106, 253–276.
- Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., et al. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. USA* 108, 1513–1518.
- Goodstadt, L., and Ponting, C.P. (2006). Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS Comput. Biol.* 2, e133.
- Gori, F., Friedman, L.G., and Demay, M.B. (2006). Wdr5, a WD-40 protein, regulates osteoblast differentiation during embryonic bone development. *Dev. Biol.* 295, 498–506.
- Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652.
- Hedges, S.B., Dudley, J., and Kumar, S. (2006). TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* 22, 2971–2972.
- Heger, A., and Ponting, C.P. (2007). Evolutionary rate analyses of orthologs and paralogs from 12 Drosophila genomes. *Genome Res.* 17, 1837–1849.
- Heide-Jørgensen, M.P., Garde, E., Nielsen, N.H., Andersen, O.N., and Hansen, S.H. (2012). A note on biological data from the hunt of bowhead whales in West Greenland 2009–2011. *J. Cetacean Res. Manag.* 12, 329–333.
- Hoeghe, C., Pfander, B., Moldovan, G.L., Pyrowolakis, G., and Jentsch, S. (2002). RAD6-dependent DNA repair is linked to modification of PCNA by ubiquitin and SUMO. *Nature* 419, 135–141.
- Holland, P.W., Garcia-Fernández, J., Williams, N.A., and Sidow, A. (1994). Gene duplications and the origins of vertebrate development. *Dev. Suppl.* 125–133.
- Holt, C., and Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 12, 491.
- Huang, W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57.
- Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. *Genome Res.* 20, 1313–1326.
- Kayagaki, N., Phung, Q., Chan, S., Chaudhari, R., Quan, C., O'Rourke, K.M., Eby, M., Pietras, E., Cheng, G., Bazan, J.F., et al. (2007). DUBA: a deubiquitinase that regulates type I interferon production. *Science* 318, 1628–1632.
- Keane, M., Craig, T., Alföldi, J., Berlin, A.M., Johnson, J., Seluanov, A., Gorbunova, V., Di Palma, F., Lindblad-Toh, K., Church, G.M., and de Magalhães, J.P. (2014). The Naked Mole Rat Genome Resource: facilitating analyses of cancer and longevity-related adaptations. *Bioinformatics* 30, 3558–3560.
- Kim, E.B., Fang, X., Fushan, A.A., Huang, Z., Lobanov, A.V., Han, L., Marino, S.M., Sun, X., Turanov, A.A., Yang, P., et al. (2011). Genome sequencing reveals insights into physiology and longevity of the naked mole rat. *Nature* 479, 223–227.
- Kurihara, L.J., Kikuchi, T., Wada, K., and Tilghman, S.M. (2001). Loss of Uchl-1 and Uchl-3 leads to neurodegeneration, posterior paralysis and dysphagia. *Hum. Mol. Genet.* 10, 1963–1970.
- Lee, D.Y., Hayes, J.J., Pruss, D., and Wolffe, A.P. (1993). A positive role for histone acetylation in transcription factor access to nucleosomal DNA. *Cell* 72, 73–84.
- Li, Y., and de Magalhães, J.P. (2013). Accelerated protein evolution analysis reveals genes and pathways associated with the evolution of mammalian longevity. *Age (Dordr.)* 35, 301–314.
- Lohse, M., Bolger, A.M., Nagel, A., Fernie, A.R., Lunn, J.E., Stitt, M., and Usadel, B. (2012). RobiNA: a user-friendly, integrated software solution for RNA-seq-based transcriptomics. *Nucleic Acids Res.* 40, W622–W627.
- López-Otín, C., Blasco, M.A., Partridge, L., Serrano, M., and Kroemer, G. (2013). The hallmarks of aging. *Cell* 153, 1194–1217.
- Löytynoja, A., and Goldman, N. (2008). A model of evolution and structure for multiple sequence alignment. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 363, 3913–3919.
- Morgan, C.C., Foster, P.G., Webb, A.E., Pisani, D., McInerney, J.O., and O'Connell, M.J. (2013). Heterogeneous models place the root of the placental mammal phylogeny. *Mol. Biol. Evol.* 30, 2145–2156.
- Oziol, E.M., Bigorgne, E., Aguilar, L., Usenko, S., and Matson, C.W. (2014). Evolved resistance to PCB- and PAH-induced cardiac teratogenesis, and reduced CYP1A activity in Gulf killifish (*Fundulus grandis*) populations from the Houston Ship Channel, Texas. *Aquat. Toxicol.* 150, 210–219.
- Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23, 1061–1067.
- Philo, L.M., Shotts, E.B., and George, J.C. (1993). Morbidity and mortality. In *The Bowhead Whale*, J.J. Burns, J.J. Montague, and C.J. Cowles, eds. (Lawrence, Kansas: Allen Press), pp. 275–312.
- Quesada, V., Ordóñez, G.R., Sánchez, L.M., Puente, X.S., and López-Otín, C. (2009). The Degradome database: mammalian proteases and diseases of proteolysis. *Nucleic Acids Res.* 37, D239–D243.
- Rogina, B., Helfand, S.L., and Frankel, S. (2002). Longevity regulation by Drosophila Rpd3 deacetylase and caloric restriction. *Science* 298, 1745.
- Ross, P.L., Cheng, I., Liu, X., Cicek, M.S., Carroll, P.R., Casey, G., and Witte, J.S. (2009). Carboxypeptidase 4 gene variants and early-onset intermediate-to-high risk prostate cancer. *BMC Cancer* 9, 69.
- Seim, I., Ma, S., Zhou, X., Geraschenko, M.V., Lee, S.G., Suydam, R., George, J.C., Bickham, J.W., and Gladyshev, V.N. (2014). The transcriptome of the bowhead whale *Balaena mysticetus* reveals adaptations of the longest-lived mammal. *Aging (Albany, N.Y. Online)* 6, 879–899.
- Tacutu, R., Craig, T., Budovsky, A., Wuttke, D., Lehmann, G., Taranukha, D., Costa, J., Fraielfeld, V.E., and de Magalhães, J.P. (2013). Human Ageing

- Genomic Resources: integrated databases and tools for the biology and genetics of ageing. *Nucleic Acids Res.* **41**, D1027–D1033.
- Tanno, M., Ogiwara, M., and Taguchi, T. (1996). Age-related changes in proliferating cell nuclear antigen levels. *Mech. Ageing Dev.* **92**, 53–66.
- Tervo, O.M., Christoffersen, M.F., Parks, S.E., Kristensen, R.M., and Madsen, P.T. (2011). Evidence for simultaneous sound production in the bowhead whale (*Balaena mysticetus*). *J. Acoust. Soc. Am.* **130**, 2257–2262.
- Tian, X., Azpurua, J., Hine, C., Vaidya, A., Myakishev-Rempel, M., Abulaeva, J., Mao, Z., Nevo, E., Gorbunova, V., and Seluanov, A. (2013). High-molecular-mass hyaluronan mediates the cancer resistance of the naked mole rat. *Nature* **499**, 346–349.
- van Beekum, O., Gao, Y., Berger, R., Koppen, A., and Kalkhoven, E. (2012). A novel RNAi lethality rescue screen to identify regulators of adipogenesis. *PLoS ONE* **7**, e37680.
- van Dam, S., Craig, T., and de Magalhães, J.P. (2015). GeneFriends: a human RNA-seq-based gene and transcript co-expression database. *Nucleic Acids Res.* <http://dx.doi.org/10.1093/nar/gku1042>
- Vervoort, V.S., Viljoen, D., Smart, R., Suthers, G., DuPont, B.R., Abbott, A., and Schwartz, C.E. (2002). Sorting nexin 3 (SNX3) is disrupted in a patient with a translocation t(6;13)(q21;q12) and microcephaly, microphthalmia, ectrodactyly, prognathism (MMEP) phenotype. *J. Med. Genet.* **39**, 893–899.
- Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* **19**, 327–335.
- Vindeløv, L.L., and Christensen, I.J. (1994). Detergent and proteolytic enzyme-based techniques for nuclear isolation and DNA content analysis. *Methods Cell Biol.* **41**, 219–229.
- Vinogradov, A.E. (1998). Genome size and GC-percent in vertebrates as determined by flow cytometry: the triangular relationship. *Cytometry* **31**, 100–109.
- Weeda, G., Donker, I., de Wit, J., Morreau, H., Janssens, R., Vissers, C.J., Nigg, A., van Steeg, H., Bootsma, D., and Hoeijmakers, J.H.J. (1997). Disruption of mouse ERCC1 results in a novel repair syndrome with growth failure, nuclear abnormalities and senescence. *Curr. Biol.* **7**, 427–439.
- West, G.B., Woodruff, W.H., and Brown, J.H. (2002). Allometric scaling of metabolic rate from molecules and mitochondria to cells and mammals. *Proc. Natl. Acad. Sci. USA* **99** (Suppl 1), 2473–2478.
- Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591.
- Yang, Z., and dos Reis, M. (2011). Statistical properties of the branch-site test of positive selection. *Mol. Biol. Evol.* **28**, 1217–1228.
- Yang, Z., Wong, W.S.W., and Nielsen, R. (2005). Bayes empirical bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.* **22**, 1107–1118.
- Yim, H.S., Cho, Y.S., Guang, X., Kang, S.G., Jeong, J.Y., Cha, S.S., Oh, H.M., Lee, J.H., Yang, E.C., Kwon, K.K., et al. (2014). Minke whale genome and aquatic adaptation in cetaceans. *Nat. Genet.* **46**, 88–92.
- Zhang, J., Zhang, Y.P., and Rosenberg, H.F. (2002). Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nat. Genet.* **30**, 411–415.
- Zhang, J., Nielsen, R., and Yang, Z. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* **22**, 2472–2479.

### **3. The Naked Mole Rat Genome Resource: facilitating analyses of cancer and longevity-related adaptations**

#### **Introduction**

The naked mole rat (NMR) is of great research interest due to its exceptional longevity and resistance to cancer. Although the NMR genome sequence has previously been published, a higher quality assembly (sequenced by the Broad Institute) and annotation was subsequently generated but not analysed or published. We therefore completed an analysis of this annotation to investigate if there were further genes with evidence of positive selection that had not been identified in the published study. We also created an online web portal from which all the data and results of our analysis can be freely downloaded.

For this work, I downloaded the assembly and annotation, calculated the dN/dS ratios, completed the analysis of positive selection, formatted the results to be made available on the portal and wrote the paper. When considering the dN/dS ratios as provided on the portal, it should be borne in mind that a value exceeding 1 is not necessarily indicative of positive selection as it could simply be due to chance fluctuations in dN and/or dS. Therefore account would need to be taken of both the p-value and the substantial number of statistical tests employed before positive selection could be predicted with confidence.



# The Naked Mole Rat Genome Resource: facilitating analyses of cancer and longevity-related adaptations

Michael Keane<sup>1,†</sup>, Thomas Craig<sup>1,†</sup>, Jessica Alföldi<sup>2</sup>, Aaron M. Berlin<sup>2</sup>, Jeremy Johnson<sup>2</sup>, Andrei Seluanov<sup>3</sup>, Vera Gorbunova<sup>3</sup>, Federica Di Palma<sup>2,4</sup>, Kerstin Lindblad-Toh<sup>2,5</sup>, George M. Church<sup>6</sup> and João Pedro de Magalhães<sup>1,\*</sup>

<sup>1</sup>Integrative Genomics of Ageing Group, Institute of Integrative Biology, University of Liverpool, Liverpool, UK, <sup>2</sup>Broad Institute of MIT and Harvard, Cambridge, MA, USA, <sup>3</sup>Department of Biology, University of Rochester, NY, USA, <sup>4</sup>Vertebrate and Health Genomics, The Genome Analysis Center, Norwich, UK, <sup>5</sup>Department of Medical Biochemistry and Microbiology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden and <sup>6</sup>Department of Genetics, Harvard Medical School, Boston, MA, USA

Associate Editor: John Hancock

## ABSTRACT

**Motivation:** The naked mole rat (*Heterocephalus glaber*) is an exceptionally long-lived and cancer-resistant rodent native to East Africa. Although its genome was previously sequenced, here we report a new assembly sequenced by us with substantially higher N50 values for scaffolds and contigs.

**Results:** We analyzed the annotation of this new improved assembly and identified candidate genomic adaptations which may have contributed to the evolution of the naked mole rat's extraordinary traits, including in regions of p53, and the hyaluronan receptors CD44 and HMMR (RHAMM). Furthermore, we developed a freely available web portal, the Naked Mole Rat Genome Resource (<http://www.naked-mole-rat.org>), featuring the data and results of our analysis, to assist researchers interested in the genome and genes of the naked mole rat, and also to facilitate further studies on this fascinating species.

**Availability and implementation:** The Naked Mole Rat Genome Resource is freely available online at <http://www.naked-mole-rat.org>. This resource is open source and the source code is available at <https://github.com/maglab/naked-mole-rat-portal>.

**Contact:** [jp@senescence.info](mailto:jp@senescence.info)

Received on April 18, 2014; revised on August 8, 2014; accepted on August 21, 2014

## 1 INTRODUCTION

The naked mole rat (NMR; *Heterocephalus glaber*) is a long-lived subterranean rodent native to the Horn of Africa. It can not only live to >30 years, making it the longest-lived rodent, but is also extremely resistant to neoplasia (Buffenstein, 2008; Tian *et al.*, 2013), and as a result is an ideal model for research on longevity, cancer and disease resistance. The NMR genome was sequenced at the BGI in 2011 to 92-fold coverage with a contig N50 of 19.3 kb and scaffold N50 of 1.6 Mb (Kim *et al.*, 2011). Here we describe a higher quality assembly (HetGla\_female\_1.0), which has subsequently been sequenced by us at the Broad Institute,

its analysis and availability on a purpose-built portal, the Naked Mole Rat Genome Resource (<http://www.naked-mole-rat.org>).

## 2 METHODS

Briefly, high molecular weight DNA was extracted from tissues of a single partially inbred female adult NMR obtained from the colony established by Vera Gorbunova at the University of Rochester, USA. The founder animals originated from the colony of J.U. Jarvis, at the University of Cape Town, South Africa. The *Heterocephalus glaber* assembly, HetGla\_female\_1.0, was constructed from 180 bp paired end fragment libraries (45 × coverage), 3 kb jumping libraries (42 × coverage), 6–14 kb sheared jumping libraries (2 × coverage) and 40 kb FOSILLs (Williams *et al.*, 2012) (1 × coverage). All libraries were sequenced by Hi-Seq Illumina machines, producing 101 bp paired-end reads. Assembly of the NMR genome was carried out using the software program ALLPATHS-LG (Gnerre *et al.*, 2011) version R38830 with default parameters.

## 3 RESULTS

HetGla\_female\_1.0 has substantially higher N50 for contigs (47.8 kb) and scaffolds (20.5 Mb) when compared with the Kim *et al.* assembly (Table 1). NG50 values, based on a C-value of 2.9 pg (source: [http://www.genomesize.com/result\\_species.php?id=4474](http://www.genomesize.com/result_species.php?id=4474)), are also considerably higher for HetGla\_female\_1.0: 35.3 kb for contigs (versus 18.1 kb for the Kim *et al.* assembly) and 20.0 Mb for scaffolds (versus 1.5 Mb).

To assist researchers in studying the genome and genes of the NMR to improve understanding of its extraordinary traits, and also to foster further studies employing this fascinating species, we developed a freely available web portal, the Naked Mole Rat Genome Resource (<http://www.naked-mole-rat.org>). Our portal features an annotation of the HetGla\_female\_1.0 assembly generated by the NCBI using the NCBI Eukaryotic Genome Annotation Pipeline (<http://www.ncbi.nlm.nih.gov/books/NBK169439/>). To assess the accuracy of this annotation, 4578 proteins were identified which exhibit at least 99% length conservation between human, mouse, rat and guinea pig orthologs. Of these, 3413 exceed the same 99% length threshold using

\*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

**Table 1.** Global statistics of the HetGla\_female\_1.0 (alias: hetGla2) assembly in comparison with the Kim *et al.* assembly (HetGla\_1.0)

Statistics	hetGla2	HetGla_1.0
RefSeq Assembly ID	GCF_000247695.1	GCF_000230445.1
Total sequence length	2 618 204 639	2 643 961 837
Number of scaffolds	4229	39 266
Scaffold N50	20 532 749	1 603 177
Number of contigs	114 653	273 990
Contig N50	47 778	21 750

the annotation of the HetGla\_female\_1.0 assembly, compared with 2158 using the annotation of Kim *et al.*

All annotated NMR sequences derived from the NCBI annotation of HetGla\_female\_1.0 are available on our portal: 42 117 coding sequences, 1779 non-coding sequences and 41 963 proteins. The 12 837 best-match NMR transcripts were identified based on coding sequence length similarity with the guinea pig ortholog, for which protein alignments and Ka/Ks ratios, calculated with the CodeML program of the PAML package v3.14 (Yang, 2007) using default parameters and guinea pig, mouse, rat and human orthologs, are also included on the portal. Genes that have been associated with longevity are cross-linked with the GenAge database (<http://www.naked-mole-rat.org/annotations/results/genage/>) (Tacutu *et al.*, 2013). A BLAST interface is also provided to allow users to quickly and easily search for sequences of interest (including coding and non-coding sequences, proteins and scaffolds). We have previously also sequenced the NMR transcriptome, which allowed us to compare liver gene expression profiles between NMRs and wild-derived mice (Yu *et al.*, 2011). The data and results of this work can also be downloaded ([http://www.naked-mole-rat.org/static/downloads/RNA\\_seq\\_supplements.zip](http://www.naked-mole-rat.org/static/downloads/RNA_seq_supplements.zip)). Moreover, an additional ~23-fold coverage assembly of the NMR genome generated by The Genome Analysis Centre (TGAC) based on two Illumina paired-end sequencing runs is available for download ([http://www.naked-mole-rat.org/static/downloads/naked\\_mole\\_rat\\_contigs.zip](http://www.naked-mole-rat.org/static/downloads/naked_mole_rat_contigs.zip)).

Guinea pig genes were used to analyse NMR orthologs of potential significance because it is the most closely related species with a high coverage genome. Functionally enriched DAVID (v6.7) clusters (using human/mouse orthologs and a background of the 12 837 best-match transcripts; otherwise default parameters were used) with an enrichment score >1.3, corresponding to  $P < 0.05$  (Huang *et al.*, 2009), for the top 5% of NMR genes by Ka/Ks included cytokine activity, signal peptide and defense response (Table 2).

Given the higher quality of this more recent genome annotation, we assessed whether we could identify novel candidate genes in the NMR that were not detected by Kim *et al.* In particular, because p53 substitutions identical to those found in human tumours have been identified in the related blind mole rat *Spalax ehrenbergi* (Ashur-Fabian *et al.*, 2004), it is relevant to assess whether there is any evidence of adaptive evolution in NMR p53. While the NMR p53 coding sequence is, not surprisingly, subject to purifying selection (Ka/Ks = 0.26), a window

**Table 2.** DAVID clusters of the highest-ranked genes by Ka/Ks between NMR and guinea pig

Cluster	Enrich. Score	No. genes	No. annots.
Signal peptide	11.86	165	10
Cytokine	10.53	36	4
Defense response	9.87	50	3
Immunoglobulin domain	6.35	34	7
Cell surface	6.19	28	3

from codons 41–80 was observed, encompassing transactivation domain 2 (TADII) and most of the proline-rich domain (PRD), which had a signature of positive selection (Ka/Ks = 2.19). The PRD is found between residues 58–98 and 55–95 of the human and mouse proteins, respectively (Walker and Levine, 1996). The human PRD contains numerous prolines including five PXXP (P = proline, X = any amino acid) motifs, compared with only two in mouse and one in rat (Toledo *et al.*, 2007). Interestingly, the NMR PRD substitutions include four proline residues, resulting in an additional four PXXP motifs relative to the guinea pig domain (Fig. 1).

This raises the possibility of convergent evolution of additional prolines and PXXP motifs in the p53 PRDs of humans and NMRs, two species which evolved an extended lifespan and consequent requirement for an enhanced DNA damage response. In addition, there are two NMR substitutions in the 9aaTADII, which has been reported to mediate apoptosis by activating targets, including *MDM2* and *BAX* (Zhu *et al.*, 1998).

Numerous proteins have been shown to interact with p53, including *BRCA1* via a region from residues 224–500 (Zhang *et al.*, 1998). There is a strong signal of selection within this region of NMR *BRCA1*, particularly from codons 430–470 (Ka/Ks > 10), which may influence the interaction with p53.

Early contact inhibition (ECI) has been identified as a novel anti-cancer mechanism in the NMR (Seluanov *et al.*, 2009), with high-molecular mass hyaluronan as the extracellular signal, which is partly transmitted via the *CD44* receptor (Tian *et al.*, 2013). Interestingly, a signal of selection (Ka/Ks > 1) was observed not only in *CD44*, from guinea pig codons 401–440, 501–540 and 661–700, but also in another hyaluronan receptor, *HMMR* (*RHAMM*), from codons 321–360, 381–420 and 441–480, suggesting that it may also contribute to transmission of the ECI signal.

Kim *et al.* reported that relative to mice, two early stop codons in the NMR p16<sup>Ink4a</sup> transcript were predicted to produce a truncated protein. There are no *Cdkn2a* transcripts in the NCBI annotation; however, a predicted transcript was generated based on alignments of the mouse and guinea pig exons with the assembly and transcriptome. Although there are no significant differences with the transcript predicted by Kim *et al.*, it is important to note that the guinea pig protein is also of similar length and shorter than in mice, indicating that this is not an NMR-specific adaptation (Fig. 2).

In conclusion, we have developed a NMR portal using a genome assembly of superior quality for the research community to benefit from this data. Our portal is designed so it can be easily updated if the NMR genome annotation is updated in

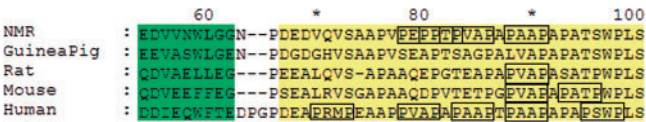


Fig. 1. Alignment of p53 sequences from NMR, guinea pig, rat, mouse and human. The TADII is in green, the PRD in yellow and PXXP motifs are boxed

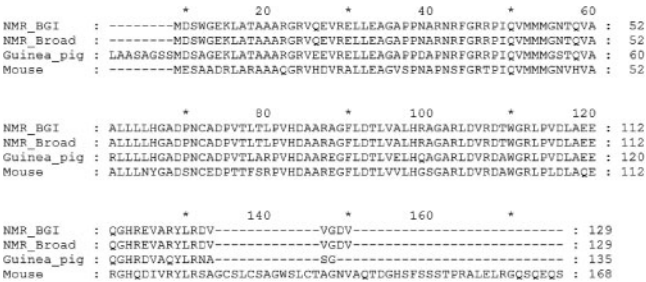


Fig. 2. Alignment of p16 sequences from guinea pig, mouse and NMR using both the Broad and BGI assemblies

the future. We also performed a reanalysis of the NMR genome using this improved assembly, which revealed further candidate genes of potential relevance to adaptive changes in the context of aging and cancer. We hope this research will facilitate and encourage studies in these amazing animals.

ACKNOWLEDGEMENTS

We would like to thank the Genomics Platform of the Broad Institute for sequencing the naked mole rat genome. Further thanks to the NCBI for the annotation of the genome and to Susan Hiatt for her assistance. We are also thankful to TGAC and the BBSRC for the generation of additional genomic data.

Funding: This work was partly funded by a Marie Curie International Reintegration Grant within EC-FP7 to J.P.M., Senior Scholar grants from the Ellison Medical Foundation to V.G. and G.M.C. and a US National Institutes of Health grant to V.G. Genome sequencing and assembly of the naked mole rat by the Broad Institute of MIT and Harvard was supported by

grants from the National Human Genome Research Institute (NHGRI). T.C. is supported by a Wellcome Trust grant (WT094386MA) to J.P.M. and M.K. is supported by a student-ship from the University of Liverpool's Faculty of Health and Life Sciences.

Conflict of Interest: none declared.

REFERENCES

Ashur-Fabian,O. et al. (2004) Evolution of p53 in hypoxia-stressed Spalax mimics human tumor mutation. *Proc. Natl Acad. Sci. USA*, **101**, 12236–12241.

Buffenstein,R. (2008) Negligible senescence in the longest living rodent, the naked mole-rat: insights from a successfully aging species. *J. Comp. Physiol. B.*, **178**, 439–445.

Gnerre,S. et al. (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl Acad. Sci. USA*, **108**, 1513–1518.

Huang,D.W. et al. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.

Kim,E.B. et al. (2011) Genome sequencing reveals insights into physiology and longevity of the naked mole rat. *Nature*, **479**, 223–227.

Seluanov,A. et al. (2009) Hypersensitivity to contact inhibition provides a clue to cancer resistance of naked mole-rat. *Proc. Natl Acad. Sci. USA*, **106**, 19352–19357.

Tacutu,R. et al. (2013) Human ageing genomic resources: integrated databases and tools for the biology and genetics of ageing. *Nucleic Acids Res.*, **41**, D1027–D1033.

Tian,X. et al. (2013) High-molecular-mass hyaluronan mediates the cancer resistance of the naked mole rat. *Nature*, **499**, 346–349.

Toledo,F. et al. (2007) Mouse mutants reveal that putative protein interaction sites in the p53 proline-rich domain are dispensable for tumor suppression. *Mol. Cell. Biol.*, **27**, 1425.

Walker,K.K. and Levine,A.J. (1996) Identification of a novel p53 functional domain that is necessary for efficient growth suppression. *Proc. Natl Acad. Sci. USA*, **93**, 15335–15340.

Williams,L.J. et al. (2012) Paired-end sequencing of Fosmid libraries by Illumina. *Genome Res.*, **22**, 2241–2249.

Yang,Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, **24**, 1586–1591.

Yu,C. et al. (2011) RNA sequencing reveals differential expression of mitochondrial and oxidation reduction genes in the long-lived naked mole-rat (*Heterocephalus glaber*) when compared to mice. *PLoS One*, **6**, e26729.

Zhang,H. et al. (1998) BRCA1 physically associates with p53 and stimulates its transcriptional activity. *Oncogene*, **16**, 1713–1721.

Zhu,J. et al. (1998) Identification of a novel p53 functional domain that is necessary for mediating apoptosis. *J. Biol. Chem.*, **273**, 13030–13036.



#### **4. MYCN/LIN28B/Let-7/HMGA2 pathway implicated by meta-analysis of GWAS in suppression of post-natal proliferation thereby potentially contributing to aging**

##### **Introduction**

A notable feature of eutherians is that their progression through the life cycle is highly synchronised with relative variation only in timing, for example the aging pathologies of old mice are largely equivalent to those of old humans and other well-studied eutherian species (Finch, 1990). In addition, there is a strong and robust positive correlation in mammals between lifespan and time to reproductive maturity, and negative correlation between lifespan and growth rate (de Magalhães et al., 2007), suggesting the possibility of a genetic timing mechanism influencing the entire life history. In support of this hypothesis, it was found that in high-density conditions, red squirrels grow significantly faster and have a reduced adult life-span (Dantzer et al., 2013). Because intra-specific trait variation is a prerequisite for the subsequent evolution of inter-specific adaptation, genomic loci associated with this hypothesised mechanism should in theory be identifiable by meta-analysing GWAS of human life history and developmental traits.

For this work, I conceived the study, downloaded the data, completed the analysis and wrote the paper. The GWAS data was obtained from the NHGRI GWAS Catalog, which is a quality controlled, manually curated, literature-derived collection of all published genome-wide association studies assaying at least 100,000 SNPs and all SNP-trait associations with p-values  $< 1.0 \times 10^{-5}$  (Hindorff et al., 2009).



## Short communication

**MYCN/LIN28B/Let-7/HMGA2 pathway implicated by meta-analysis of GWAS in suppression of post-natal proliferation thereby potentially contributing to aging**

Michael Keane, João Pedro de Magalhães \*

Integrative Genomics of Ageing Group, Institute of Integrative Biology, University of Liverpool, Liverpool L69 7ZB, UK

## ARTICLE INFO

## Article history:

Received 7 December 2012  
 Received in revised form 12 April 2013  
 Accepted 19 April 2013  
 Available online 29 April 2013

## Keywords:

HMGA2  
 LIN28B  
 Suppression of proliferation  
 Pace of development  
 Aging

## ABSTRACT

Mammalian organ and body growth slows and finally terminates because of a progressive suppression of cell proliferation, however little is known about the genetic regulatory mechanisms responsible. A meta-analysis of genome-wide association studies using growth and development-related traits revealed that two genes, *HMGA2* and *LIN28B*, had multiple associations. Altered *HMGA2* expression has been shown to result in both overgrowth and pygmy phenotypes in mice and overgrowth in humans. These genes are members of the *MYCN/LIN28B/Let-7/HMGA2* pathway and homologs of *LIN28B* and *let-7* are known to regulate developmental timing in *Caenorhabditis elegans*. Strikingly, expression levels of *let-7* and *Hmga2* in murine stem cells continue to increase and decrease, respectively, after growth terminates, suggesting that this pathway may contribute to regulating the pace of both development and age-related degenerative phenotypes.

© 2013 Elsevier Ireland Ltd. All rights reserved.

Mammalian somatic growth progressively slows postnatally primarily due to a progressive decline in cell proliferation (Chang et al., 2008), however the genetic regulatory mechanisms responsible remain largely obscure (Kennedy and Norman, 2005). Consistent with the antagonistic pleiotropy theory (Williams, 1957), recent results suggest the existence of a multi-organ genetic program suppressing proliferation (PSP) which progressively down-regulates many growth-promoting genes (Lui et al., 2010a) and persists into adulthood, thereby potentially contributing to aging (Lui et al., 2010b).

To identify potential regulatory components of the PSP, a meta-analysis of genome-wide association studies (GWAS) from the National Human Genome Research Institute GWAS catalog (Hindorff et al., 2009) was performed. The 9217 SNPs in the GWAS catalog were filtered for growth and development-related traits resulting in a dataset with 428 SNPs from 45 studies associated with 11 traits. Genes reported in the studies as associated with the SNPs were employed.

Permutation testing is commonly used to determine significance (Johnson et al., 2010) and was employed to estimate the false-discovery rate. In each of 10,000 iterations, all SNPs were randomly and independently assigned to the estimated 22,333 human protein-coding genes retrieved from NCBI Entrez (Pruitt et al., 2009), and the gene with the maximum number of SNPs was

identified. Genes with more than two SNPs occurred in 2.43% of the iterations, establishing genes with three or more SNPs as significant. Sixteen reported genes reached this threshold and those with the most associations with multiple developmental traits were *HMGA2* and *LIN28B* with 14 and 7 associated SNPs, respectively (Table 1). In an additional control analysis using 428 randomly selected SNPs, a gene with more than six SNPs was observed in 1.9% of 10,000 iterations, confirming the significance of these two genes.

*HMGA2* is a member of the high-mobility group A family that can modulate transcription by altering chromatin structure (Reeves, 2001). Supporting the validity of the association with postnatal proliferation and growth is the case of an individual with a chromosomal inversion truncating this gene, resulting in slightly elevated expression of 1.4 times that of a control (Ligon et al., 2005). Notable phenotypes at 8 years of age were extreme overgrowth in terms of height, weight and head circumference, advanced bone age (~13.5 years) and arthritis. In addition, this individual developed premature dentition, and a panoramic dental X-ray at 4 years indicated advanced dental age. Similarly, expression of a truncated *Hmga2* induced gigantism in transgenic mice (Battista et al., 1999). By contrast, *Hmga2*-null mice demonstrate the “pygmy” phenotype characterized by dramatic reductions in body fat and small stature (Zhou et al., 1995).

*LIN28B* is a homolog of the *Caenorhabditis elegans* *lin-28* gene (Guo et al., 2006) which controls developmental timing (Moss et al., 1997). *LIN28B* negatively regulates *let-7* (Piskounova et al., 2011) which in turn is a negative regulator of *HMGA2* (Lee and

\* Corresponding author.

E-mail addresses: [jp@senescence.info](mailto:jp@senescence.info), [aging@liverpool.ac.uk](mailto:aging@liverpool.ac.uk) (J.P. de Magalhães).

**Table 1**

The developmental traits used in the meta-analysis and results for the top two genes. The remaining statistically significant genes are: *ADAMTSL3*, *CDK6*, *DLEU7*, *DYM*, *EFEMP1*, *GNA12*, *GPR126*, *HHIP*, *HMGAI*, *LCORL*, *LTBP1*, *MSRB3*, *PLAG1* and *ZBTB38*.

Trait	<i>HMGAI</i>			<i>LIN28B</i>		
	Number of SNPs	Number of studies	Context(s)	Number of SNPs	Number of studies	Context(s)
Height	8	8	UTR-3 (6) Intergenic Intron	2	2	Intron (2)
Head circumference (infant)	1	1	UTR-3	–		
Brain structure	1	1	Intron	–		
Normalized brain volume	–			–		
Hippocampal volume	–			–		
Primary tooth development (number of teeth)	1	1	Intergenic	–		
Primary tooth development (time to first tooth eruption)	1	1	Intergenic	–		
Permanent tooth development	1	1	Intergenic	–		
Aortic root size	1	1	Intergenic	–		
Menarche (age at onset)	–			4	4	Intron (2) Intergenic (2)
Menarche and menopause (age at onset)	–			1	1	Intron

Dutta, 2007). *Let-7* was the second microRNA discovered and has also been shown to regulate developmental timing in *C. elegans* (Reinhart et al., 2000). Furthermore, *let-7* and the *DAF-12* nuclear hormone receptor engage in reciprocal direct feedback regulation (Hammell et al., 2009), and it was recently shown that upon induction by *DAF-12*, *let-7* can stimulate *DAF-16/FOXO* signaling to extend life by targeting *lin-14* and *akt-1* (Shen et al., 2012). This is particularly significant because of the impact of *daf-12* on insulin/IGF-1 signaling (Cypser et al., 2006), which plays a well-established regulatory role in both development and aging (Cohen and Dillin, 2008). In humans, *LIN28B* and *HMGAI* are members of the oncogenic *MYCN/LIN28B/Let-7/HMGAI* pathway (Helland et al., 2011).

It was recently observed that there are dramatic changes in the expression of the *Lin28/let-7* axis in the rat hypothalamus during postnatal maturation (Sangiao-Alvarellos et al., 2013), and *LIN28B* over-expression was also shown to increase *MYCN* levels and induce neuroblastoma by suppressing *let-7* (Molenaar et al., 2012). *Mycn* was identified as a transcription factor that is consistently down-regulated during development in multiple mouse and rat organs (Lui et al., 2010a). Similarly, *Hmgai* expression is significantly higher in fetal than young-adult stem cells (Kiel et al., 2005), and it is required to maintain stem cell self-renewal in multiple tissues (Nishino et al., 2008). Furthermore, *Hmgai* levels inversely correlate with expression of *let-7* (Mayr et al., 2007). While it is possible that additional mechanisms influence phenotypes, perhaps via early life effects, *Hmgai* was not found to be required for stem cell formation during embryonic development (Nishino et al., 2008).

These findings indicate that the *MYCN/LIN28B/Let-7/HMGAI* pathway may be an important regulatory component of the PSP.

Because body size is maintained following growth termination, it might be expected that *let-7* and *Hmgai* levels in stem cells would stabilize. Therefore it is particularly notable that, to the contrary, they continue to increase and decrease in expression, respectively, coinciding with increasing expression of *p16<sup>lnk4a</sup>*, a potent tumor suppressor (Nishino et al., 2008). *p16<sup>lnk4a</sup>* and *p19<sup>Arf</sup>* levels in stem cells are negatively regulated by *Hmgai* (Nishino et al., 2008) and over-expression of *p16<sup>lnk4a</sup>* with age has been reported to decrease stem cell self-renewal in mice (Molofsky et al., 2006). This suggests that the PSP continues to progress into adulthood, which has been hypothesized to contribute to aging (Lui et al., 2010b). Further supporting this hypothesis is gene expression data showing that many of the changes that occur during aging originate during development and that cell-cycle-related genes are strongly over-represented among genes that

persistently decline in expression throughout postnatal life (Lui et al., 2010b). In addition, increasing expression of *let-7* has been shown to contribute to declining germ-line stem cell self-renewal in *Drosophila* (Toledano et al., 2012) and human neurodegeneration (Lehmann et al., 2012). A QTL encompassing *Hmgai* has also been associated with longevity in mice (Klebanov et al., 2001).

Because continuation of the PSP after growth terminates will ultimately cause deleterious degenerative phenotypes, it could be assumed that it would have been strongly selected against. One possibility is that it escaped further selective pressure once manifestation of these phenotypes was delayed until the end of the typical reproductive lifespan (de Magalhães, 2012), during which it might also have a fitness-enhancing effect by slightly reducing cancer risk and energy requirement. It could also be expected that a steady decrease in body size would be observed due simply to net cell loss, which clearly conflicts with reality. Conversely it has been demonstrated that senescent cells accumulate in mammalian tissue from early adulthood (Herbig et al., 2006). However not enough is currently known to support firm conclusions about mechanisms maintaining organ and body size.

Taken together, these results link this pathway to a growth-regulation process potentially relevant to aging, hence it merits further studies. *Hmgai*-null mice have been proposed as a model to test if cell divisions contribute to aging (de Magalhães and Faragher, 2008). While their increased suppression of stem cell proliferation could in isolation be expected to decrease lifespan, a probable confounding factor is their small body size which likely results in reduced demand on stem cell pools. Indeed, these opposing effects suggest a possible explanation for the puzzling observation that the correlation of longevity with body size is negative intra-species but positive inter-species (Miller et al., 2002). In larger species such as humans relative to mice, the greatly increased chronological delay in the induction of *p16<sup>lnk4a</sup>* (Kim and Sharpless, 2006) suggests that the rate of change in expression of its regulators *HMGAI* and *let-7* has correspondingly been significantly reduced, with a positive effect on longevity. However if this discount rate of stem cell self-renewal is intra-specifically consistent, it appears plausible that smaller individuals would experience a slower rate of tissue degeneration due simply to a lower total cellularity representing a reduced cell replacement burden on stem cell pools. Therefore while the net impact on the lifespan of *Hmgai*-null mice is difficult to predict, it seems improbable that no longevity effects would be observed. The timing of growth termination in these animals relative to wild type may indicate which effect is dominant. An alternative experiment would be to maintain expression of one or more members of this

pathway at growth termination levels, perhaps using transgenic *Drosophila* and in particular the known orthologs *lin-28* and *let-7*. It could reasonably be anticipated that a modest increase in energy requirement and hyperplasia together with a significant attenuation of age-related degenerative phenotypes would be observed.

## Acknowledgements

The authors would like to thank Daniel Wuttke for critical reading of the manuscript. MK is supported by a studentship from the University of Liverpool's Faculty of Health and Life Sciences.

## References

- Battista, S., Fidanza, V., Fedele, M., Klein-Szanto, J.P., Outwater, E., Brunner, H., Santoro, M., Croce, C.M., Fusco, A., 1999. The expression of a truncated HMGIC gene induces gigantism associated with lipomatosis. *Cancer Research* 59, 4793–4797.
- Chang, M., Parker, E.A., Muller, T.J., Haenen, C., Mistry, M., Finkelstein, G.P., Murphy-Ryan, M., Barnes, K.M., Sundaram, R., Baron, J., 2008. Changes in cell-cycle kinetics responsible for limiting somatic growth in mice. *Pediatric Research* 64, 240–245.
- Cohen, E., Dillin, A., 2008. The insulin paradox: aging, proteotoxicity and neurodegeneration. *Nature Reviews. Neuroscience* 9, 759–767.
- Cypser, J.R., Tedesco, P., Johnson, T.E., 2006. Hormesis and aging in *Caenorhabditis elegans*. *Experimental Gerontology* 41, 935–939.
- de Magalhães, J.P., 2012. Programmatic features of aging originating in development: aging mechanisms beyond molecular damage? *FASEB Journal* 26, 4821–4826.
- de Magalhães, J.P., Faragher, R.G.A., 2008. Cell divisions and mammalian aging: integrative biology insights from genes that regulate longevity. *BioEssays* 30, 567–578.
- Guo, Y., Chen, Y., Ito, H., Watanabe, A., Ge, X., Kodama, T., Aburatani, H., 2006. Identification and characterization of *lin-28* homolog B (*LIN28B*) in human hepatocellular carcinoma. *Gene* 384, 51–61.
- Hammell, C.M., Karp, X., Ambros, V., 2009. A feedback circuit involving *let-7*-family miRNAs and *DAF-12* integrates environmental signals and developmental timing in *Caenorhabditis elegans*. *Proceedings of the National Academy of Sciences of the United States of America* 106, 18668–18673.
- Helland, Å., Anglesio, M.S., George, J., Cowin, P.A., Johnstone, C.N., House, C.M., et al., 2011. Deregulation of *MYCN*, *LIN28B* and *LET7* in a molecular subtype of aggressive high-grade serous ovarian cancers. *PLoS ONE* 6, e18064.
- Herbig, U., Ferreira, M., Condel, L., Carey, D., Sedivy, J.M., 2006. Cellular senescence in aging primates. *Science* 311, 1257.
- Hindorf, L.A., Sethupathy, P., Jenkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., Manolio, T.A., 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America* 106, 9362–9367.
- Johnson, R.C., Nelson, G.W., Troyer, J.L., Lautenberger, J.A., Kessing, B.D., Winkler, C.A., O'Brien, S.J., 2010. Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC Genomics* 11, 724.
- Kennedy, D., Norman, C., 2005. What don't we know? *Science* 309, 75.
- Kiel, M.J., Iwashita, T., Yilmaz, O.H., Morrison, S.J., 2005. Spatial differences in hematopoiesis but not in stem cells indicate a lack of regional patterning in definitive hematopoietic stem cells. *Developmental Biology* 283, 29–39.
- Kim, W.Y., Sharpless, N.E., 2006. The regulation of *INK4/ARF* in cancer and aging. *Cell* 127, 265–275.
- Klebanov, S., Astle, C.M., Roderick, T.H., Flurkey, K., Archer, J.R., et al., 2001. Maximum life spans in mice are extended by wild strain alleles. *Experimental Biology and Medicine* (Maywood, NJ) 226, 854–859.
- Lee, Y.S., Dutta, A., 2007. The tumor suppressor microRNA *let-7* represses the *HMG2* oncogene. *Genes & Development* 21, 1025–1030.
- Lehmann, S.M., Kruger, C., Park, B., Derkow, K., Rosenberger, K., Baumgart, J., et al., 2012. An unconventional role for miRNA: *let-7* activates Toll-like receptor 7 and causes neurodegeneration. *Nature Neuroscience* 15, 827–835.
- Ligon, A.H., Moore, S.D.P., Parisi, M.A., Mealiffe, M.E., Harris, D.J., Ferguson, H.L., Quade, B.J., Morton, C.C., 2005. Constitutional rearrangement of the architectural factor *HMG2*: a novel human phenotype including overgrowth and lipomas. *American Journal of Human Genetics* 76, 340–348.
- Lui, J.C., Forcinito, P., Chang, M., Chen, W., Barnes, K.M., Baron, J., 2010a. Coordinated postnatal down-regulation of multiple growth-promoting genes: evidence for a genetic program limiting organ growth. *FASEB Journal* 24, 3083–3092.
- Lui, J.C., Chen, W., Barnes, K.M., Baron, J., 2010b. Changes in gene expression associated with aging commonly originate during juvenile growth. *Mechanisms of Ageing and Development* 131, 641–649.
- Mayr, C., Hemann, M.T., Bartel, D., 2007. Disrupting the pairing between *let-7* and *HMG2* enhances oncogenic transformation. *Science* 315, 1576–1579.
- Miller, R.A., Harper, J.M., Galecki, A., Burke, D.T., 2002. Big mice die young: early life body weight predicts longevity in genetically heterogeneous mice. *Aging Cell* 1 (1) 22–29.
- Molenaar, J.J., Domingo-Fernández, R., Ebus, M.E., Lindner, S., Koster, J., et al., 2012. *LIN28B* induces neuroblastoma and enhances *MYCN* levels via *let-7* suppression. *Nature Genetics* 44, 1199–1206.
- Molofsky, A.V., Slutsky, S.G., Joseph, N.M., He, S., Pardal, R., Krishnamurthy, J., Sharpless, N.E., Morrison, S.J., 2006. Increasing *p16INK4a* expression decreases forebrain progenitors and neurogenesis during ageing. *Nature* 443, 448–452.
- Moss, E.G., Lee, R.C., Ambros, V., 1997. The cold shock domain protein *LIN-28* controls developmental timing in *C. elegans* and is regulated by the *lin-4* RNA. *Cell* 88, 637–646.
- Nishino, J., Kim, I., Chada, K., Morrison, S.J., 2008. *Hmg2* promotes neural stem cell self-renewal in young but not old mice by reducing *p16INK4a* and *p19Arf* expression. *Cell* 135, 227–239.
- Piskounova, E., Polyarchou, C., Thornton, J.E., Lapierre, R.J., Pothoulakis, C., Hagan, J.P., Iliopoulos, D., Gregory, R.I., 2011. *LIN28A* and *LIN28B* inhibit *let-7* microRNA biogenesis by distinct mechanisms. *Cell* 147, 1066–1079.
- Pruitt, K.D., Tatusova, T., Klimke, W., Maglott, D.R., 2009. NCBI reference sequences: current status, policy and new initiatives. *Nucleic Acids Research* 37 (Database) D32–D36.
- Reeves, R., 2001. Molecular biology of *HMG*A proteins: hubs of nuclear function. *Gene* 277, 63–81.
- Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., Rougvie, A.E., Horvitz, H.R., Ruvkun, G., 2000. The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature* 403, 901–906.
- Sangiao-Alvarellos, S., Manfredi-Lozano, M., Ruiz-Pino, F., Navarro, V.M., Sánchez-Garrido, M.A., Leon, S., Dieguez, C., Cordido, F., Matagne, V., Dissen, G.A., Ojeda, S.R., Pinilla, L., Tena-Sempere, M., 2013. Changes in hypothalamic expression of the *LIN28/let-7* system and related microRNAs during postnatal maturation and after experimental manipulations of puberty. *Endocrinology* 154 (2) 942–955.
- Shen, Y., Wollam, J., Magner, D., Karalay, O., Antebi, A., 2012. A steroid receptor-microRNA switch regulates life span in response to signals from the gonad. *Science* 338, 1472–1476.
- Toledano, H., D'Alterio, C., Czech, B., Levine, E., Jones, D.L., 2012. The *let-7*-*Imp* axis regulates ageing of the *Drosophila* testis stem-cell niche. *Nature* 485, 605–610.
- Williams, G.C., 1957. Pleiotropy, natural selection, and the evolution of senescence. *Evolution* 11, 398–411.
- Zhou, X., Benson, K.F., Ashar, H.R., Chada, K., 1995. Mutation responsible for the mouse pygmy phenotype in the developmentally regulated factor *HMGIC*. *Nature* 377, 771–774.

## 5. Systematic detection of positive selection on human 3'UTRs

### Introduction

Although the most significant human adaptations include those related to skeletal morphology, encephalisation and longevity, their molecular basis remains largely unknown (Varki et al., 2008). Genes associated with these traits do not feature prominently in surveys of human coding sequences for evidence of positive selection, which have primarily identified genes associated with immune response, sensory perception and reproduction (Haygood et al., 2010). Although one study reported accelerated evolution of a set of human nervous system-related genes (Dorus et al., 2004), a subsequent analysis of a larger set of brain-specific genes found no evidence of acceleration on the human lineage, suggesting that the results of the previous study were based on a biased dataset (Shi et al., 2006).

Recent results support the proposition that the genomic basis of adaptive evolution lies primarily in non-coding regulatory elements rather than coding sequences (Carroll, 2005). In relation to analysis of human data, over 80% of GWAS loci have non-coding variants as the likely causal association (Hindorff et al., 2009). In addition, a multiple-population genome-wide study of marine-to-freshwater adaptation in stickleback fish reported that only 17% of loci bearing the genomic hallmarks of adaptive evolution were protein coding (Jones, 2012), with a similar estimate of 20% in mice (Halligan et al., 2013). Finally, a genome-wide survey found that 92% of human accelerated regions were non-coding (Lindblad-Toh et al., 2011). These results correlate well with a recent study which reported that approximately 8.2% of the human genome is presently subject to negative selection (Rands et al., 2014), of which over 80% must be non-coding. However, analysis of selection on regulatory sequences presents formidable difficulties relative to coding sequences, not least the lack of functional annotation. For example, although there are estimated to be hundreds of thousands of enhancers in the human genome, only a tiny fraction have been identified and functionally assayed to date (ENCODE Project Consortium, 2012). In addition, the target genes of regulatory elements are often not located nearby (Lettice et al., 2003) or even on the same chromosome (Spilianakis et al., 2005). Finally, there is not yet a consensus on the appropriate neutral rates to use when assessing substitution rates on non-coding sequences (Zhen and Andolfatto, 2012).

However, one category of non-coding sequences that are well annotated in the human genome is untranslated regions (UTRs). In particular, 3'UTRs typically contain multiple binding sites for

regulatory proteins and microRNAs, and interestingly an analysis of these regions from a small set of human brain-expressed genes found no evidence of accelerated evolution (Li and Su, 2006). In contrast, it was shown that a significant proportion (~ 60%) of the nucleotide divergence in *Drosophila* UTRs was driven to fixation by positive selection (Andolfatto, 2005). However, the extent of selection on human 3'UTRs does not appear to have been systematically assessed to date, therefore to address this issue a genome-wide survey of these regions was performed.

## Methods

The following data were retrieved from Ensembl ([www.ensembl.org](http://www.ensembl.org)):

- Complete DNA sequences of the following mammalian genomes: *Homo sapiens*, *Pan troglodytes*, *Gorilla gorilla* and *Mus musculus*.
- Chimpanzee, gorilla and mouse orthologs for all human protein-coding genes.
- Genomic co-ordinates of all human 3'UTRs.

The mouse and gorilla genomes were included in order to allow identification of conserved regions and assignment of substitutions to the human or chimpanzee lineages respectively. All human genes with a 1:1 ortholog in both chimpanzee and gorilla were identified, and the longest 3'UTR starting after the final codon was analysed. The human sequence was aligned with the orthologous chimpanzee, gorilla and mouse chromosomes using the BLAST (Altschul et al, 1990) blastn program, which was used in order to exclude regions that align poorly.

To identify elements of functional significance, the aligned chimpanzee, gorilla and mouse sequences were used to identify regions of greater than 90% identity over more than 30 base pairs. For each of these regions, the human-specific substitutions, defined as sites where the human base differed from the identical chimpanzee and gorilla base, were then identified. Analogous to the commonly used dN/dS test for protein coding sequences (The Chimpanzee Sequencing and Analysis Consortium, 2005), regions in which the human substitution rate exceeded the local intergenic/intronic substitution rate (The Chimpanzee Sequencing and Analysis Consortium, 2005) were identified as putatively positively selected.

In addition, an empirical  $p$ -value was calculated for each region as the proportion of 10,000 randomly-generated regions of identical length with a substitution rate equalling the local intergenic/intronic rate which had a greater or equal concentration of human-specific substitutions. For example, to test a region with a length of 30 bases and 3 substitutions, a sequence of 300,000 bases was constructed and substitutions were randomly assigned until the substitution rate equalled the relevant local neutral rate. Each successive 30 base region in this sequence was then assessed to obtain the number of regions in which the number of substitutions was at least 3, which allowed the calculation of the empirical  $p$ -value. Benjamini-Hochberg correction was used to control the false-discovery rate (FDR) using an FDR of 0.1.

GC-biased gene conversion (gBGC) is a recombination-associated process that favours the fixation of G/C alleles over A/T alleles (W→S), and can contribute to a spurious signal of selection (Duret and Galtier, 2009). To assess the probability of gBGC, an empirical  $p$ -value was computed for every

human 3'UTR as the proportion of 10,000 instances of the orthologous chimpanzee sequence with an equivalent number of randomly-generated human-specific substitutions which had an equal or higher proportion of W→S substitutions. All human W→S substitutions were excluded for 3'UTRs with  $p < 0.05$ .

Due to evidence that the proportion of simultaneous double-nucleotide substitutions in primate genomes is significant (Averof et al., 2000), all tandem-base substitutions were regarded as a single mutation event.

The Database for Annotation, Visualization and Integrated Discovery (DAVID) (Huang et al., 2009) was used to identify functional groups with an enrichment score exceeding 1.3, corresponding to  $p < 0.05$  (Huang et al., 2009), with a background of all genes analysed.



## Results and discussion

A total of 33,864 conserved elements were identified in the 3'UTRs of 13,567 human genes, and the human substitution rates in 145 of these elements were found to exceed the local neutral rate. The corresponding genes exhibit a functional enrichment for neuron differentiation and limb development (Table 5.1).

Table 5.1: top DAVID clusters of human genes with evidence of 3'UTR positive selection.

Cluster	Enrich. score	No. of annot.	No. of genes
Neuron differentiation, developmental protein	2.92	14	12
Limb development, embryonic morphogenesis	2.08	22	10
Ion binding	1.88	8	46
Pattern specification process	1.57	3	8

However as a result of the large number of statistical tests employed, none of these regions survived Benjamini-Hochberg FDR correction at the 0.1 level. It is also important to note that neuron-expressing transcripts tend to have longer 3' UTRs (Zhang et al., 2005) and thus are more likely to contain conserved elements than randomly selected genes. As such, although the regions identified cannot be regarded with confidence as statistically significant, the list ranked in order of increasing p-value does at the very least provide a set of regions that may be worthy of further investigation.

The reason that the method described here did not identify many conserved regions exhibiting significant evidence of accelerated evolution may be due at least in part to the minimum length of 30 base pairs used. In particular, microRNAs are a vital component of gene control and primarily exert their effects by targeting short (average length of 8 bp) target sequences in 3' UTRs (Xie et al., 2005). A method which cannot detect conserved elements shorter than 30 bp will clearly miss selection on such short sequences.

In addition, it may be necessary to employ genomes of more closely related species to human than mouse in order to more accurately identify conserved sequence and improve statistical power. In

particular, phylogenetic shadowing (Boffelli et al., 2003), where sequences from a large set of closely-related species are compared in order to identify conserved regulatory elements, may be a suitable method to improve the sensitivity of an analysis of selection on non-coding regions.

## 6. Discussion and conclusions

The results presented herein were obtained from comparative analysis of annotated coding sequences of bowhead whale and naked mole rat (NMR), and human data related to genome wide association studies (GWAS) and 3' UTRs.

The analysis of bowhead whale genes identified several previously associated with cancer and ageing exhibiting evidence of positive selection. Gene duplications related to DNA repair, cell cycle regulation, cancer and ageing were also observed, and in addition alterations in genes linked to thermoregulation, sensory perception, dietary adaptations and immune response were identified. However there are some limitations that must be acknowledged. Firstly, significant issues were observed with the annotation that was generated for publication, which dramatically reduced the number of bowhead genes that could be analysed. Although the MAKER2 annotation which I subsequently generated seems to be superior, there are still significant issues which were observed with the results. This serves to illustrate the challenges inherent in the use of automated computational annotation pipelines, a fact which is also apparent in relation to issues with numerous annotated cow genes. In addition, the statistical significance of the comparative analyses of selection and the number of statistical tests used was not accounted for and as such, the results cannot be regarded as high-confidence instances of positive selection as they could also simply be neutral changes or the consequence of genetic drift.

Evidence of positive selection was found in several NMR genes of interest which were not previously reported, including in regions of p53 and the hyaluronan receptors *CD44* and *HMMR*. However it should be noted that a multiple testing correction was not applied, so these results should not be regarded as high-confidence. As such, future work should include assessment of the statistical significance of the dN/dS results in addition to accounting for the false discovery rate. Another suggestion would be to systematically assess the quality of the gene annotations in order to allow subsequent predictions of selection to be made with greater confidence.

The meta-analysis of GWAS data found evidence that a conserved pathway regulated by *Let-7* and *LIN-28B* is involved in the timing of multiple human developmental events. Intriguingly, *LIN-28* and *let-7* have been characterised in *C. elegans* and are known to regulate developmental progression (Ambros, 2011). The data used in this meta-analysis was obtained from the NHGRI GWAS Catalog, which is a quality controlled, manually curated, literature-derived collection of all published genome-wide association studies assaying at least 100,000 SNPs and all SNP-trait associations with

p-values  $< 1.0 \times 10^{-5}$  (Hindorff et al., 2009). While there are several issues which can potentially confound the results of a GWAS analysis, linkage disequilibrium perhaps being the most obvious, it should be noted that numerous quality-control steps are included in the preparation of the GWAS catalog which greatly mitigate against this risk. Firstly, data extraction and curation for the GWAS Catalog is an expert activity; each step is performed manually by scientists supported by a web-based tracking and data entry system which allows multiple curators to search, annotate, verify and publish the Catalog data. In addition, only the most significant SNP from each independent locus is included and for each SNP, an associated gene is only included if gene information was found and reported by the author(s). In addition, it is also confirmed that the reported SNP and gene are in the same chromosomal location.

It has recently been demonstrated that convergent phenotypic evolution can be the result of genomic and molecular convergence (Parker et al., 2013). In this context, it is interesting that evidence of adaptive evolution has been observed in *Uncoupling protein 1 (UCP1)* from both naked mole rat (Kim et al., 2011) and bowhead whale (Keane et al., 2014), which may be a result of selection on thermoregulation in both species. Furthermore, this raises the intriguing question of the extent of convergent molecular evolution in these two species related to their exceptional cancer-resistance and longevity, perhaps their most complex and interesting adaptations.

Although the mechanisms underlying bowhead anti-cancer adaptations are currently unknown, significant progress has been made towards understanding the cancer-resistance of the NMR. In particular, it was shown that a unique sequence of *Hyaluronan synthase 2 (HAS2)* contributed to the mediation of early contact inhibition in this species (Tian et al., 2013). Interestingly, from the analyses of positive selection and unique residues that are available for NMR and bowhead whale, the only common gene is *Apex nuclease 1 (APEX1)*, a DNA repair enzyme.

Disruption of *APEX1* in mice results in embryonic lethality (Xanthoudakis et al., 1996), and heterozygous mice exhibit an elevated spontaneous mutation frequency (Huamani et al., 2004). Unfortunately there is only a single unique residue in bowhead, therefore this cannot be regarded as a high-confidence result. The annotated bowhead gene also contained two indels which were found to be spurious following manual annotation. This did not result in the identification of any further unique residues, however.

It is somewhat disappointing that there are not additional genes with evidence of molecular adaptation in both species. There are a number of potential explanations for this observation, the first and most obvious being that different genes and pathways were targeted during the evolution

of enhanced cancer-resistance and longevity in these species. Alternatively, if either sequencing or annotation is not of sufficiently high quality, a large proportion of genes may have to be filtered from the analysis in order to minimise false positives, thus severely limiting the power and value of the analysis. Accurately annotated coding sequences are clearly essential in particular to assess the evidence for positive selection, as otherwise incorrectly annotated regions can be mistaken for selection, resulting in a very high false positive rate. This is unfortunately of relevance in relation to the bowhead annotation, as many genes were found upon manual inspection to have suspicious indels, which were concluded to be spurious in the cases of the minority that were subsequently manually annotated.

In addition, it is important to note that the bowhead and NMR genome analyses have not extended beyond coding sequences, as is typical for genome sequencing projects. Indeed, molecular evolution research has focused almost exclusively on coding sequences, for which there are several reasons. Firstly, coding sequences are well annotated and synonymous sites provide an obvious estimate of the local neutral mutation rate with which the substitution rate in non-synonymous sites can be compared. By contrast, functional non-coding sequences are extremely difficult to identify, and even then it is typically unclear which genes they are targeting. However in spite of the numerous analyses of primate coding sequences that have been reported to date, the results in relation to human evolution remain speculative, with the notable exceptions of *FOXP2* and *CMAH* (Varki et al., 2008). This suggests that the major contributor to phenotypic divergence may be adaptation in regulatory sequences, and indeed recent results from analyses of GWAS (Hindorff et al., 2009), sticklebacks (Jones et al., 2012), mice (Halligan et al., 2013) and humans (Lindblad-Toh et al., 2011) consistently estimate that over 80% of adaptive molecular evolution occurs in non-coding DNA. In addition, there is evidence that the more complex and arguably interesting traits, such as neural development and function, are due to adaptation in non-coding rather than coding sequences (Haygood et al., 2010). There are significant obstacles in the analysis of selection on gene regulation however, with the almost complete lack of annotation of non-coding regions among the most significant. In addition, perhaps the single most serious issue with putative cases of positive selection identified to date is the lack of subsequent functional characterisation (Vallender, 2012), which renders the results speculative.

In view of these issues, analysis and characterisation of proximal non-coding DNA, that is UTRs and promoters, may provide a potential way forward. The genes regulated by these regions are evident, which allows assay of the effect of molecular adaptations by transfection and surprisingly, the extent of positive selection does not appear to have been systematically assessed to date.

In conclusion, the work described herein identified a number of cases of genomic adaptations which may be relevant in terms of trait adaptation. The analyses of bowhead whale, NMR and human genomic data each found evidence of divergence in genes that have been associated with variation in longevity, cancer-resistance or life history timing. In addition, the data generated have been made available online in order to facilitate further research on these species.

However, there are also some limitations that must be acknowledged. Firstly, the quality of the annotation for the bowhead analysis in particular would need to be improved to increase the number of genes which can be analysed with confidence. In addition, as with the many solely bioinformatic analyses of molecular evolution that have been reported at this point, without experimental validation the findings must be regarded as speculative. It is envisaged that in order to significantly advance the field, such functional characterisation must become a routine part of projects analysing molecular evolution, in order to avoid simply adding to an ever-increasing collection of speculative results.

## References

- Alkan, C., Sajjadian, S., Eichler, E.E., 2011. Limitations of next-generation genome sequence assembly. *Nat. Methods* 8: 61–65.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215: 403-410.
- Ambros, V, 2011. MicroRNAs and developmental timing. *Curr. Opin. Genet. Dev.* 21 (4): 511-7.
- Andolfatto, P., 2005. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* 437 (7062): 1149-52.
- Anisimova, M., Nielsen, R., et al., 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* 164 (3): 1229-1236.
- Assefa, S., Keane, T. M., Otto, T. D., Newbold, C., Berriman, M., 2009. ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics* 25: 1968–1969.
- Austad, S.N., 2010. Methusaleh’s Zoo: how nature provides us with clues for extending human health span. *J. Comp. Pathol.* 142 (Suppl. 1): S10–S21.
- Austad, S. N., Fischer, K. E., 1991. Mammalian aging, metabolism, and ecology: evidence from the bats and marsupials. *J. Gerontol.* 46 (2): B47-53.
- Averof, M., Rokas, A., Wolfe, K.H., Sharp, P.M., 2000. Evidence for a High Frequency of Simultaneous Double-Nucleotide Substitutions. *Science* 287 (5456): 1283-1286.
- Bairoch, A., Boeckmann, B., Ferro, S., Gasteiger, E., 2004. Swiss-Prot: juggling between evolution and stability. *Brief. Bioinform.* 5: 39–55.
- Bakewell, M.A., Shi, P., Zhang, J., 2007. More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proc. Natl. Acad. Sci. USA* 104: 7489-7494.
- Banerjee, A.K., 1980. 5'-terminal cap structure in eukaryotic messenger ribonucleic acids. *Microbiol. Rev.* 44: 175–205.
- Barrett, L.W., Fletcher, S., Wilton, S.D., 2012. Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. *Cell Mol. Life Sci.* 69 (21): 3613–3634.
- Bejerano, G., Pheasant, M., et al., 2004. Ultraconserved elements in the human genome. *Science* 304: 1321–1325.

- Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I., Pachter, L., Rubin, E.M., 2003. Phylogenetic Shadowing of Primate Sequences to Find Functional Regions of the Human Genome. *Science* 28: 1391-1394.
- Cantarel, B. L. et al., 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* 18: 188–196.
- Carroll, S.B., 2005. Evolution at Two Levels: On Genes and Form. *PLoS Biology* 3 (7): e245.
- Caulin, A. F., Maley, C. C., 2011. Peto's Paradox: evolution's prescription for cancer prevention. *Trends in ecology & evolution* 26: 175-182.
- Cho, Y.S., et al., 2013. The tiger genome and comparative analysis with lion and snow leopard genomes. *Nature Communications* 4 (2433): 10.1038.
- Clark, A.G., et al., 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* 302: 1960-1963.
- The Chimpanzee Sequencing and Analysis Consortium, 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437: 69-87.
- Dantzer, B., Newman, A.E., Boonstra, R., Palme, R., Boutin, S., Humphries, M.M., McAdam, A.G., 2013. Density triggers maternal hormones that increase adaptive offspring growth in a wild mammal. *Science* 340 (6137): 1215-7.
- Darwin, C., 1859. *On the Origin of Species*.
- de Magalhães, J.P., 2013. How ageing processes influence cancer. *Nat. Rev. Cancer* 13: 357–365.
- de Magalhães, J.P., Costa, J., Church, G.M., 2007. An analysis of the relationship between metabolism, developmental schedules, and longevity using phylogenetic independent contrasts. *J. Gerontol. A. Biol. Sci. Med. Sci.* 62 (2): 149-160.
- de Magalhães, J.P., Costa, J., Toussaint, O., 2005. HAGR: the Human Ageing Genomic Resources. *Nucleic Acids Res.* 33 (Database Issue): D537–D543.
- de Magalhães, J.P., Keane, M., 2013. Endless paces of degeneration— applying comparative genomics to study evolution's moulding of longevity. *EMBO Rep.* 14: 661–662.
- Dobzhansky, T., Hecht, M.K., Steere, W.C., 1968. On some fundamental concepts of evolutionary biology. *Evolutionary Biology* 2: 1-34.



- Dorus, S., Vallender, E.J., Evans, P.D., Anderson, J.R., Gilbert, S.L., Mahowald, M., Wyckoff, G.J., Malcom, C.M., Lahn, B.T., 2004. Accelerated evolution of nervous system genes in the origin of *Homo sapiens*. *Cell* 119 (7): 1027-40.
- Duret, L., Galtier, N., 2009. Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. *Annu. Rev. Genomics Hum. Genet.* 10: 285-311.
- Eisenberg, D.T., et al., 2010. Worldwide allele frequencies of the human apolipoprotein E gene: climate, local adaptations, and evolutionary history. *Am. J. Phys. Anthropol.* 143: 100–111.
- ENCODE Project Consortium, 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489: 57–74.
- Finch, C., 1990. *Longevity, Senescence, and the Genome* (Chicago: University of Chicago Press).
- Finch, C.E., Stanford, C.B., 2004. Meat-adaptive genes and the evolution of slower aging in humans. *Q. Rev. Biol.* 79: 3–50.
- Flicek, P., Ahmed, I., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., et al., 2013. Ensembl 2013. *Nucleic Acids Res.* 41: D48–D55.
- Foote, A.D., et al., 2015. Convergent evolution of the genomes of marine mammals. *Nature Genetics* 47: 272–275.
- Fullerton, S.M., et al., 2000. Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. *Am. J. Hum. Genet.* 67: 881–900.
- Galtier, N. et al., 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics* 159: 907–911.
- Garde, E., Heide-Jorgensen, M.P., Hansen, S.H., Nachman, G., Forchhammer, M.C., 2007. Age specific growth and remarkable longevity in narwhals (*Monodon monoceros*) from West Greenland as estimated by aspartic acid racemization. *J. Mammal* 88: 49-58.
- Gatesy, J., Geisler, J.H., Chang, J., Buell, C., Berta, A., Meredith, R.W., Springer, M.S., McGowen, M.R., 2013. A phylogenetic blueprint for a modern whale. *Mol. Phylogenet. Evol.* 66: 479–506.
- George, J.C., Bada, J., Zeh, J., Scott, L., Brown, S.E., O’Hara, T., Suydam, R., 1999. Age and growth estimates of bowhead whales (*Balaena mysticetus*) via aspartic acid racemization. *Can. J. Zool.* 77: 571–580.

- Gibbs, R.A., et al., 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science* 316: 222-234.
- Gillet, L.C.J., Scharer, O.D., 2006. Molecular mechanisms of mammalian global genome nucleotide excision repair. *Chem. Rev.* 106: 253–276.
- Gnerre, S., et al., 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. USA* 108: 1513–1518.
- Gori, F., Friedman, L.G., and Demay, M.B., 2006. Wdr5, a WD-40 protein, regulates osteoblast differentiation during embryonic bone development. *Dev. Biol.* 295: 498–506.
- Halligan, D.L., et al., 2013. Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid rodents. *PLoS Genet.* 9 (12): e1003995.
- Hardy, G. H., 1908. Mendelian proportions in a mixed population. *Science* 28: 49–50.
- Haygood, R., Babbitt, C.C., Fedrigo, O., Wray, G.A., 2010. Contrasts between adaptive coding and noncoding changes during human evolution. *Proc. Natl. Acad. Sci. USA* 107 (17): 7853-7.
- Hesketh, J., 2004. 3'-Untranslated regions are important in mRNA localization and translation: lessons from selenium and metallothionein. *Biochem. Soc. Trans.* 32: 990-3.
- Hindorff, L. A. et al., 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* 106: 9362–9367.
- Holland, P. W., Garcia-Fernandez, J., Williams, N. A., Sidow, A., 1994. Gene duplications and the origins of vertebrate development. *Development*: 125-133.
- Holt, C., Yandell, M., 2011. MAKER2: an annotation pipeline and genome database management tool for second-generation genome projects. *BMC Bioinformatics* 12: 491.
- Huamani et al., 2004. Spontaneous mutagenesis is enhanced in Apex heterozygous mice. *Mol. Cell. Biol.* 24 (18): 8145-8153.
- Huang, D.W., Sherman, B.T., Lempicki, R.A., 2009. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc.* 4 (1): 44-57.
- Hunter, D. J., Altshuler, D., Rader, D. J., 2008. From Darwin's finches to canaries in the coal mine--mining the genome for new biology. *The New England Journal of Medicine* 358: 2760-3.

Husemann, P., Stoye, J., 2010. r2cat: syntenic plots and comparative assembly. *Bioinformatics* 26: 570–571.

Jabbari, K., Bernardi, G., 2004. Body temperature and evolutionary genomics of vertebrates: a lesson from the genomes of *Takifugu rubripes* and *Tetraodon nigroviridis*. *Gene* 333: 179–181.

Jirimitu et al., 2012. Genome sequences of wild and domestic bactrian camels. *Nature Communications* 3 (1202): 10.1038.

Jobson, R.W., Nabholz, B., Galtier, N., 2010. An evolutionary genome scan for longevity-related natural selection in mammals. *Mol Biol Evol.* 27(4): 840-7.

Jones, F.C., et al., 2012. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* 484: 55–61.

Keane, M., Craig, T., Alfoldi, J., Berlin, A.M., Johnson, J., Seluanov, A., Gorbunova, V., Di Palma, F., Lindblad-Toh, K., Church, G.M., de Magalhães, J.P., 2014. The Naked Mole Rat Genome Resource: facilitating analyses of cancer and longevity-related adaptations. *Bioinformatics* 30: 3558–3560.

Keane, M., et al., 2015. Insights into the evolution of longevity from the bowhead whale genome. *Cell Rep.* 10 (1): 112-22.

Kim, E. B., et al., 2011. Genome sequencing reveals insights into physiology and longevity of the naked mole rat. *Nature* 479: 223-227.

Kimura, M., 1968. Evolutionary Rate at the Molecular Level. *Nature* 217: 624-6.

Kudla, G. et al., 2006. High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biol.* 4: e180.

Kuida, K., Haydar, T.F., Kuan, C.Y., Yang, D., Karasuyama, H., Rakic, P., Flavell, R.A., 1998. Reduced apoptosis and cytochrome C-mediated caspase activation in mice lacking caspase 9. *Cell* 94: 325–337.

Lander, E. S., et al., 2001. Initial sequencing and analysis of the human genome. *Nature* 409: 860–921.

Lee, D.Y., Hayes, J.J., Pruss, D., Wolffe, A.P., 1993. A positive role for histone acetylation in transcription factor access to nucleosomal DNA. *Cell* 72: 73–84.

Lettice, L. A., et al., 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* 12: 1725–1735.

Li, Y., de Magalhães, J.P., 2013. Accelerated protein evolution analysis reveals genes and pathways associated with the evolution of mammalian longevity. *Age (Dordr)* 35: 301–314.

Li, Y., Su, B., 2006. No accelerated evolution of 3'UTR region in human for brain-expressed genes. *Gene* 383: 38–42.

Lindblad-Toh, K., Wade, C.M., et al., 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438: 803–819.

Lindblad-Toh, K., et al., 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478: 476–482.

Luo, R., et al., 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* 1:18.

Marchini, J., Howie, B., Myers, S., McVean, G., Donnelly, P., 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* 39: 906–913.

McLean, C.Y., Reno, P.L., et al., 2011. Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* 471: 216–219.

Nagylaki, T., 1983. Evolution of a finite population under gene conversion. *Proc. Natl. Acad. Sci. U.S.A.* 80: 5941–5945.

Parker, J., et al., 2013. Genome-wide signatures of convergent evolution in echolocating mammals. *Nature* 502: 228–231.

Parra, G., Bradnam, K., Korf, I., 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23: 1061–1067.

Perry, J., Ashworth, A., 1999. Evolutionary rate of a gene affected by chromosomal position. *Curr. Biol.* 9: 987–989.

Philo, L.M., Shotts, E.B., George, J.C., 1993. Morbidity and mortality. The Bowhead Whale, J.J. Burns, J.J. Montague, and C.J. Cowles, eds. (Lawrence, Kansas: Allen Press): 275–312.

Pollard, K.S., Salama, S.R., et al., 2006. Forces shaping the fastest evolving regions in the human genome. *PLoS Genet.* 2: e168.

- Promislow, D. E., 1993. On size and survival: progress and pitfalls in the allometry of life span. *J. Gerontol.* 48 (4): B115-123.
- Pruitt, K.D., et al., 2009. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.* 19 (7): 1316-23.
- Rands, C.M., Meader, S., Ponting, C.P., Lunter, G., 2014. 8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage. *PLoS Genet.* 10 (7): e1004525.
- Reese, M. G., Guigo, R, 2006. EGASP: Introduction. *Genome Biol.* 7 (Suppl. 1): 1–3.
- Rogina, B., Helfand, S.L., Frankel, S., 2002. Longevity regulation by *Drosophila* Rpd3 deacetylase and caloric restriction. *Science* 298: 1745.
- Schneider, A., et al., 2009. Estimates of Positive Darwinian Selection Are Inflated by Errors in Sequencing, Annotation, and Alignment. *Genome Biol. Evol.* 1: 114–118.
- Sebastiani, P., et al., 2011. Retraction. *Science* 22: 404.
- Sebastiani, P., et al., 2012. Genetic Signatures of Exceptional Longevity in Humans. *PLoS One* 7(1): e29848.
- Seim, I., Ma, S., Zhou, X., Gerashchenko, M.V., Lee, S.G., Suydam, R., George, J.C., Bickham, J.W., and Gladyshev, V.N. (2014). The transcriptome of the bowhead whale *Balaena mysticetus* reveals adaptations of the longest-lived mammal. *Aging (Albany, N.Y. Online)* 6: 879–899.
- Semeiks, J., Grishin, N.V., 2012. A Method to Find Longevity-Selected Positions in the Mammalian Proteome. *PLoS ONE* 7(6): e38595.
- Shaffer, H.B., et al., 2013. The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. *Genome Biology* 14: R28.
- Shapiro, M.D., et al., 2013. Genomic Diversity and Evolution of the Head Crest in the Rock Pigeon. *Science* 339 (6123): 1063-1067.
- Shi, P., Bakewell, M.A., Zhang, J., 2006. Did brain-specific genes evolve faster in humans than in chimpanzees? *Trends Genet.* 2006 22 (11): 608-13.
- Slater, G. S., Birney, E., 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6: 31.

- Smale, S.T., Kadonaga, J.T., 2003. The RNA polymerase II core promoter. *Ann. Rev. Biochem.* 72: 449–479.
- Smit, A., Hubley, R., 2011. RepeatModeler 1.05, <http://www.repeatmasker.org>.
- Soriano, P. et al., 1987. High rate of recombination and double crossovers in the mouse pseudoautosomal region during male meiosis. *Proc. Natl. Acad. Sci. U.S.A.* 84: 7218–7220.
- Spilianakis, C.G., Lalioti, M.D., Town, T., Lee, G.R., Flavell, R.A., 2005. Interchromosomal associations between alternatively expressed loci. *Nature* 435 (7042): 637–45.
- Stearns, S. C., 1992. *The Evolution of Life Histories*. Oxford University Press, Oxford.
- Tacutu, R., Craig, T., Budovsky, A., Wuttke, D., Lehmann, G., Taranukha, D., Costa, J., Fraifeld, V.E., de Magalhães, J.P., 2013. Human Ageing Genomic Resources: integrated databases and tools for the biology and genetics of ageing. *Nucleic Acids Res.* 41, D1027–D1033.
- Tervo, O.M., Christoffersen, M.F., Parks, S.E., Kristensen, R.M., Madsen, P.T., 2011. Evidence for simultaneous sound production in the bowhead whale (*Balaena mysticetus*). *J. Acoust. Soc. Am.* 130, 2257–2262.
- Tian, X., et al., 2013. High-molecular-mass hyaluronan mediates the cancer resistance of the naked mole rat. *Nature* 499, 346–349.
- Tsai, I. J., Otto, T. D., Berriman, M., 2010. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biol.* 11: R41.
- Vallender, E.J., 2012. Genetic correlates of the evolving primate brain. *Prog. Brain Res.* 195: 27–44.
- Varki, A., Geschwind, D.H., Eichler, E.E., 2008. Explaining human uniqueness: genome interactions with environment, behaviour and culture. *Nat. Rev. Genet.* 9 (10): 749–63.
- Vervoort, V.S., Viljoen, D., Smart, R., Suthers, G., DuPont, B.R., Abbott, A., Schwartz, C.E. (2002). Sorting nexin 3 (SNX3) is disrupted in a patient with a translocation t(6;13)(q21;q12) and microcephaly, microphthalmia, ectrodactyly, prognathism (MMEP) phenotype. *J. Med. Genet.* 39, 893–899.
- Vonk, F.J., et al., 2013. The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system. *Proc. Natl. Acad. Sci. USA* 110: 20651–20656.

- Weeda, G., Donker, I., de Wit, J., Morreau, H., Janssens, R., Vissers, C.J., Nigg, A., van Steeg, H., Bootsma, D., and Hoeijmakers, J.H.J., 1997. Disruption of mouse ERCC1 results in a novel repair syndrome with growth failure, nuclear abnormalities and senescence. *Curr. Biol.* 7: 427–439.
- Weinberg, W., 1908. Ueber den Nachweis der Vererbung beim Menschen. *Jahresh. Ver. Vaterl. Naturkd. Wuerttemb.* 64: 369–382.
- Wong, P., Wiley, E., Johnson, W., Ryder, O., O’Brien, S., Haussler, D, et al., 2012. Tissue sampling methods and standards for vertebrate genomics. *GigaScience* 1: 8.
- Woolfe, A., Goodson, M., et al., 2005. Highly conserved noncoding sequences are associated with vertebrate development. *PLoS Biol.* 3: e7.
- Wright, S., 1929. The evolution of dominance. *The American Naturalist* 63 (689): 556–561.
- Xanthoudakis et al., 1996. The redox/DNA repair protein, Ref-1, is essential for early embryonic development in mice. *Proc. Natl. Acad. Sci. U.S.A.* 93 (17): 8919-8923.
- Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., Kellis, M, 2005. Systematic discovery of regulatory motifs in human promoters and 3’ UTRs by comparison of several mammals. *Nature* 434: 338–345.
- Yang, Z., Nielsen, R., 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* 19: 908-917.
- Yang, Z., Wong, W.S., Nielsen, R., 2005. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.* 22 (4): 1107-18.
- Yang, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591.
- Ye, L., et al., 2011. A vertebrate case study of the quality of assemblies derived from next-generation sequences. *Genome Biol.* 12: R31.
- Yim, H.S., Cho, Y.S., Guang, X., Kang, S.G., Jeong, J.Y., Cha, S.S., Oh, H.M., Lee, J.H., Yang, E.C., Kwon, K.K., et al., 2014. Minke whale genome and aquatic adaptation in cetaceans. *Nat. Genet.* 46, 88–92.
- Zákány, J., Kmita, M., Duboule, D., 2004. A dual role for Hox genes in limb anterior-posterior asymmetry. *Science* 304 (5677): 1669–1672.

Zhang, J., 2004. Frequent false detection of positive selection by the likelihood method with branch-site models. *Mol. Biol. Evol.* 21: 1332-1339.

Zhang, G., et al., 2013. Comparative Analysis of Bat Genomes Provides Insight into the Evolution of Flight and Immunity. *Science* 339 (6118): 456-460.

Zhang, H. B., Lee, J.Y., Tian, B., 2005. Biased alternative polyadenylation in human tissues. *Genome Biol.* 6: R100.

Zhen, Y., Andolfatto, P., 2012. Methods to detect selection on non-coding DNA. *Methods Mol. Biol.* 856: 141-59.



# **Appendices**

## **Introduction**

This section consists of a published opinion paper and the Supplementary Information from the bowhead whale paper previously described in Chapter 2.

The opinion paper presents some opportunities and challenges in attempting to apply comparative genomics to the study of longevity as the quantity of genomic data increases rapidly. My contribution was to write the paper subsequent to discussing and agreeing the most relevant topics to be included.

The Supplementary Information in this section is from our publication describing the annotation and analysis of the bowhead whale genome (Keane et al., 2015) and primarily consists of analyses contributed by collaborators which could not be included in the paper due to space limitations. As first author of the paper, I collated and prepared these contributions in a format suitable for inclusion with the paper.

---

## Endless paces of degeneration—applying comparative genomics to study evolution’s moulding of longevity

*João Pedro de Magalhães & Michael Keane*

**W**hy can mice not live more than five years and dogs not more than 30, yet bats can live over 40 years and humans over a century?

Differences in longevity between closely related species are one of the greatest mysteries in biology, and identifying the processes responsible could ultimately

presage the development of therapies against a multitude of age-related diseases. The variation in mammalian longevity must have a genomic basis, with recent

genome sequencing efforts opening up exciting opportunities to decipher it; some promising results are beginning to emerge. Analysis of two bat genomes revealed that a high proportion of genes in the DNA damage checkpoint–DNA repair pathway, including *ATM*, *TP53*, *RAD50* and *KU80*, are under selection in bats [1]. This finding is exciting because these genes have been directly associated with ageing in model systems and, therefore, it points towards a potential role for averting DNA damage in longevity assurance mechanisms; a notion dating back several decades that remains contentious. In addition, the report of a systematic scan for proteins with accelerated evolution in mammalian lineages in which longevity increased over the course of their evolution, hinted that some repair systems, such as the ubiquitin–proteasome pathway and a few proteins related to DNA damage repair, might have been selected for in long-lived lineages [2]. However, much work remains to improve the signal-to-noise ratio of this and similar methods.

With decreasing costs of sequencing, the growing number of genomes of species with diverse lifespans is expected to facilitate studies in this area. As such, we can make an increasing number of comparisons such as those described above. Put simply, if we study long-lived species and find that they share genetic adaptations—for example in DNA damage response pathways—then we might assume that those adaptations are important to increase longevity. There are major intrinsic difficulties with this type of analysis, however, that one must keep in mind. Perhaps the best illustration is that despite the dramatic phenotypic divergence between humans and chimpanzees, only a relatively small number of genetic adaptations that are probably responsible

for such divergence have thus far been identified [3]. One difficulty is that the genomic elements underlying species differences remain controversial. Possible processes include mutations in coding and non-coding sequences, gene family expansion and contraction, and copy number variation, all of which we think must be explored in the context of longevity adaptations. Whilst changes in regulatory regions might be important, standard methods are lacking for the detection of selection on functional non-coding sequences on a genome-wide scale and this, we think, is a limitation for progress in this area. Another limitation is that experimental validation of promising candidates is often extremely difficult to obtain.

Applying comparative genomics to study the evolution of longevity also has unique challenges. For one, the force of natural selection weakens with age, indicating that, although under low-hazard conditions selection favours genes and pathways conferring longevity, selective pressure for longevity is significantly less than for other traits. Furthermore, we think that the integration of additional data—for example gene expression and age-related phenotypic data—is crucial to link genotypes to phenotypes and identify physiological adaptations that are required for extended longevity. Unfortunately, such data and even the necessary samples to generate it are as yet only available for a subset of species. In our opinion, another crucial issue is the extent to which common mechanisms underlie the extension of longevity by evolution in different species. Just as rare variants contribute to missing human heritability, taxa-specific adaptations might contribute to longevity. It can be assumed that the environment—for

example, diet—of each species will influence the physiological and biochemical pathways that must be optimized to fend off ageing and age-related diseases. However, the ageing process, despite progressing at different rates, is remarkably similar across most mammals studied [4], hinting that retarding ageing might involve adaptations in similar pathways. The degree of overlap between longevity assurance mechanisms is, in our view, a crucial determinant of how much we can expect to learn about species differences in ageing in the foreseeable future. If common pathways do indeed underlie longevity evolution in multiple species, even if involving different genetic elements in different taxa, then it is reasonable to expect that they can be identified by using comparative genomics as more genomes of short- and long-lived species are sequenced. We hope to live long enough to help unravel this age-old problem.

#### CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

#### REFERENCES

1. Zhang G *et al* (2013) *Science* **339**: 456–460
2. Li Y, de Magalhães JP (2013) *Age (Dordr)* **35**: 301–314
3. O’Bleness M *et al* (2012) *Nat Rev Genet* **13**: 853–866
4. Finch CE (1990) *Longevity, Senescence, and the Genome*. University of Chicago Press

**João Pedro de Magalhães and Michael Keane are at the Integrative Genomics of Ageing Group, Institute of Integrative Biology, University of Liverpool, Liverpool, UK. E-mail: jp@senescence.info**

EMBO reports (2013) **14**, 661–662; published online 12 July 2013; doi:10.1038/embor.2013.96

# Supplementary Information

---

**Keane et al. *Insights into the evolution of longevity from the bowhead whale genome***

## Contents

Genome size estimation	p. 2
RNA sequencing in Alaskan specimens	p. 5
Repeat sequences	p. 7
Gene duplication	p. 9
Analysis of bowhead whale protease genes	p. 10

## Genome size estimation

Simple ratios, assuming a chicken genome size of  $C = 1.25$  pg, were used to convert mean fluorescence to pg of DNA. Mouse and rat tissues, which were included as an additional confirmation of genome size estimation accuracy, were within 2% and 3%, respectively, of published values (data not shown). Bowhead whale genome sizes were estimated using both chicken as a size standard, and by averaging the estimates produced from all three size standards (chicken, mouse, and rat) independently. The results from these two methods yielded estimates of 2.93 and 2.92 pg, respectively. Of particular interest was the variability in individual bowhead whale genome size estimates, an approximately 3% difference between our two samples (Figure S1). While not known during sample processing and initial analysis, bowhead #10B17, the individual with the smaller genome (2.88 pg), was a male, whereas bowhead #10B18, the individual with the larger genome (2.98 pg) was a female. This difference in genome size is entirely accounted for by the expected differences in masses of X and Y chromosomes. As is customary, the final bowhead whale genome size estimate was calculated as the average of the male and female genome sizes, 2.93 pg or 2.87 Gb (Figure S1).

This is the first cytometric-based estimate of genome size for a baleen whale. The value  $C = 2.93$  pg is the lowest value yet for a cetacean (Figure S2) and is on the low end of values for Cetartiodactyla (artiodactyls and cetaceans). The average of all mammals is  $C = 3.5$  pg, so bowheads are low for mammals. Most of the mammalian species with lower genome sizes are animal with small body size and high metabolic rates including bats, shrews and some rodents. Only toothed whales are available for comparison and thus it is not known if bowheads are atypical for baleen whales. Nevertheless it is apparent from these results that bowheads are at the low end of the scale for mammals in general.

There are two possible explanations for the relatively small genome of the bowhead whale. The first is that it could be a plesiomorphic character unchanged during the evolution and diversification of cetartiodactyls. This is possible given the fact that low genome sizes are also found in suids, camelids, giraffids, cervids and bovids, notwithstanding the fact that most cetartiodactyls have higher values (<http://www.genomesize.com/>) and the ancestral character state is not known.

The second possible explanation is that the low genome size of the bowhead is a derived, adaptive, character state that has evolved as a result of nucleotypic effects. A correlate to small genome size is not obvious but could be related to metabolic rate or gas exchange in this highly specialized diving mammal.

Significance of the genome size estimate of bowheads also relates to its genome sequence. There is a discrepancy in the genome size as measured in base pairs (one picogram = 978 megabases) with flow cytometry compared to the total sequence length in the genome sequence (Figure S1). The flow cytometric method is 20% higher than the sequence total and this is likely due to the inability of the bioinformatics methods to assemble repetitive DNA sequences. So,

the estimated genome size gives us an independent estimate of the amount of sequence not represented in the assembled genome sequence.

Additional studies of genome size are needed for baleen whales in order to determine if the bowhead is an outlier or if this group of mammals has an unexpectedly small genome size. In this way perhaps the adaptive correlates, if any exist, can be determined. In addition, it is anticipated that other baleen whales will be the subjects of genome sequences and a better understanding of the amount of DNA sequence not assembled is useful for determining the overall percent coverage of the genome sequence.

## Bowhead Whale Genome Size

- Bowhead whale genome (1C) is 2.93 pg (2.87 Gb)
- Genome coverage = 2.3 Gb
  - 20% missing, possibly repetitive DNA
- Smallest documented cetacean genome
  - Six measured cetacean genomes (five different species) are all > 3.0 pg
  - Limited cetacean data available for comparison

FL2-H (fluorescence) to 1c genome size correction based on chicken size standard

$\text{FL2-H} \times 0.01165 = 1\text{C value (pg)}$

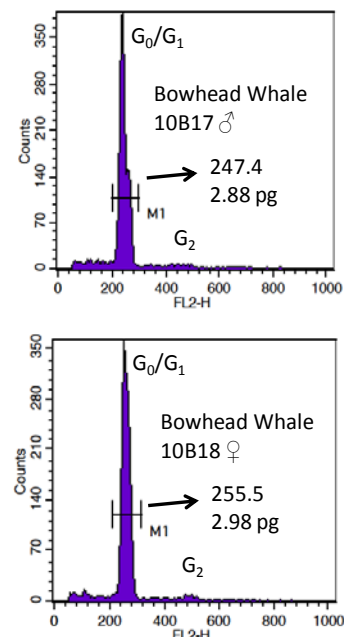
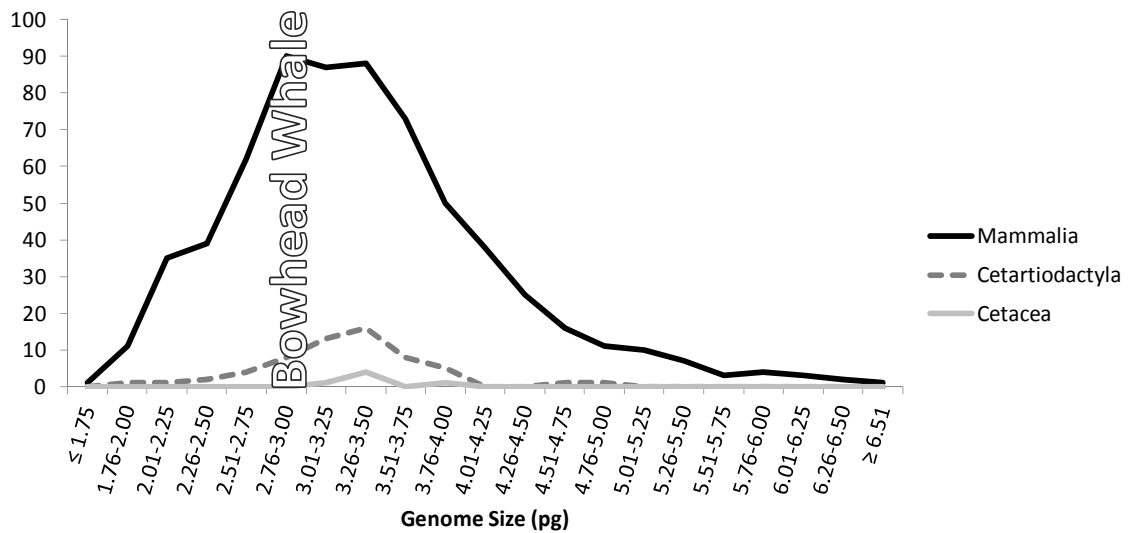


Figure S1—DNA flow histograms (right two panels) of a male and female bowhead whale showing an approximately 3% difference in estimated genome sizes. The mean estimated genome size is  $C = 2.93$  pg.

# Genome Size Distributions



Bowhead genome = 2.93 pg

Genome size data queried from the *Animal Genome Size Database*  
([www.genomesize.com](http://www.genomesize.com))

Figure S2—Distribution of genome sizes of Mammalia, Cetartiodactyla, and Cetacea. Bowhead whales have an estimated genome size (2.93 pg) well below the mammalian mean (3.5 pg). This is the first species of baleen whale to be reported and has the lowest C-value of any cetacean. Some cetartiodactyls have lower genome sizes but most are higher than bowheads.

## RNA sequencing in Alaskan specimens

Sequence analysis of RNA from 5 tissues representing two bowhead whales produced a total of 138,495,774 sequence reads comprising >13 billion bp after quality control and primer trimming. The numbers and sizes of reads and contigs are reported in Table S1. The total number of annotated contigs was 81,319. The estimated number of bowhead contigs identified as being homologous to human genes was approximately 14,000 or ca. 60% of the known human genes.

Table S1—Results of RNA sequencing of 5 tissues from two bowhead whales. All Reads refers to all sequenced fragments of any size, Large Contigs includes contigs comprised of multiple reads of 500 bp or larger, and All Contigs refers to small and large contigs combined.

<b>All Reads</b>	<b>Total reads</b>	138,495,774
	<b>Total bases</b>	13,162,565,851
	<b>Size range of reads</b>	2-101
	<b>N50 (modal size)</b>	101
	<b>Average length</b>	95
<b>Large Contigs</b>	<b>Contig size</b>	≥500
	<b>Total large contigs</b>	157,699
	<b>Total number of bases</b>	322,342,312
	<b>Contig size range</b>	500-24765
	<b>N50 (modal size)</b>	3,442
	<b>Average length</b>	2,044
<b>All Contigs</b>	<b>Total number of contigs</b>	423,657
	<b>Total number of bases</b>	401,340,157
	<b>Contig size range</b>	201-24765
	<b>N50 (modal size)</b>	2,436
	<b>Average length</b>	947
<b>Annotations</b>	<b>Number of annotated contigs</b>	81,319



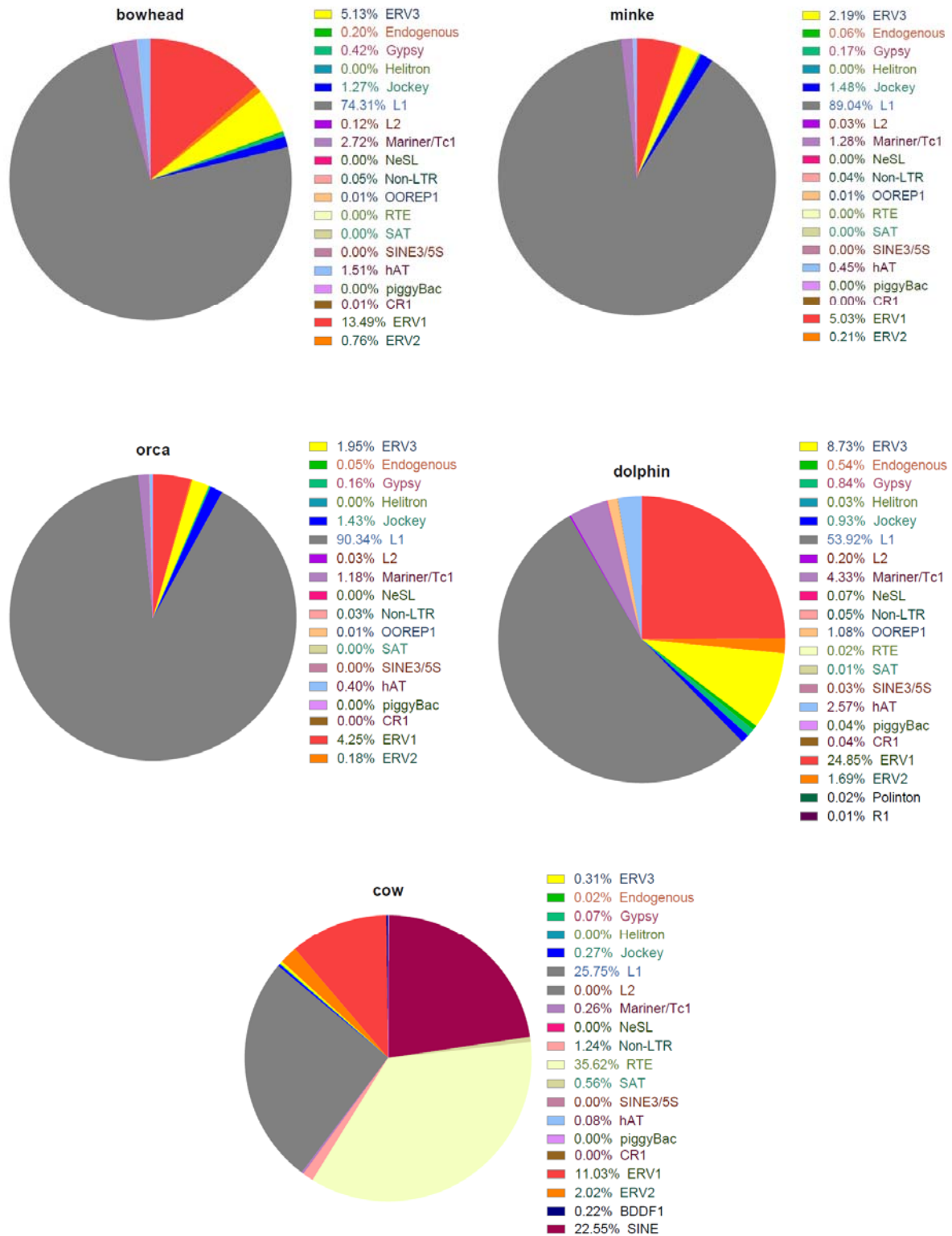
Table S2 shows the estimated frequencies of SNPs among the 5 tissues sampled. The two individuals sampled can be compared by reference to retina (bowhead 10B20) and Tissues 1-4 (bowhead 10B16). The data are shown for 8 size classes of contigs. As contigs size increases, the frequency of estimated SNPs increases. With this method, there appears to be approximately 0.5-0.6 SNPs per 1,000 bases of RNA.

Table S2—SNP frequencies estimated for each tissue per size class of contigs. Tissues 1-4 are from bowhead 10B16 and retina is from 10B20.

Contig Size (bp)	Tissue					
	1. Cerebellum	2. Heart	3. Liver	4. Testes	5. Retina	Tissues 1-4
≥201	2.7E-04	2.7E-04	2.7E-04	2.8E-04	3.1E-04	3.9E-04
>500	3.3E-04	3.2E-04	3.2E-04	3.4E-04	3.8E-04	4.8E-04
>1000	3.6E-04	3.5E-04	3.6E-04	3.8E-04	4.2E-04	5.2E-04
>2000	3.9E-04	3.8E-04	3.8E-04	4.1E-04	4.5E-04	5.6E-04
>3000	4.0E-04	3.9E-04	3.9E-04	4.2E-04	4.6E-04	5.7E-04
>4000	4.2E-04	4.0E-04	4.0E-04	4.4E-04	4.7E-04	5.9E-04
>5000	4.3E-04	4.0E-04	4.0E-04	4.5E-04	4.7E-04	6.0E-04
>6000	4.5E-04	4.2E-04	4.2E-04	4.7E-04	4.9E-04	6.2E-04

## Repeat sequences

Figure S3: Transposable elements



**Table S3. Transposable elements**

<b>#repeats</b>	<b>bowhead</b>	<b>minke</b>	<b>cow</b>	<b>orca</b>	<b>dolphin</b>
BDDF1	0	0	3849	0	0
CR1	20	18	10	17	55
ERV1	41312	55718	191377	46093	33419
ERV2	2329	2356	35066	1984	2268
ERV3	15696	24226	5309	21139	11734
Endogenous	621	642	280	568	727
Gypsy	1273	1874	1298	1718	1125
Helitron	6	5	1	3	42
Jockey	3878	16395	4610	15502	1248
L1	227486	985534	446716	980177	72518
L2	376	350	55	300	272
Mariner/Tc1	8336	14194	4452	12819	5821
NeSL	2	1	10	1	90
Non-LTR	154	426	21430	286	65
OOREP1	19	103	0	103	1459
Polinton	0	0	0	0	28
R1	0	0	0	0	11
RTE	1	1	618004	0	24
SAT	11	24	9776	6	17
SINE	0	0	391287	0	0
SINE3/5S	7	10	27	10	42
hAT	4608	4991	1349	4299	3462
piggyBac	3	6	4	4	59

## Gene duplication

**Table S4: Branch-site test Bayes empirical Bayes values for putative positively selected sites in PCNA. \*Indicates statistical significance.**

Site	Sub.	BEB
34	V	0.774
38	H	0.753
58	S	0.983*
103	L	0.758
231	T	0.748

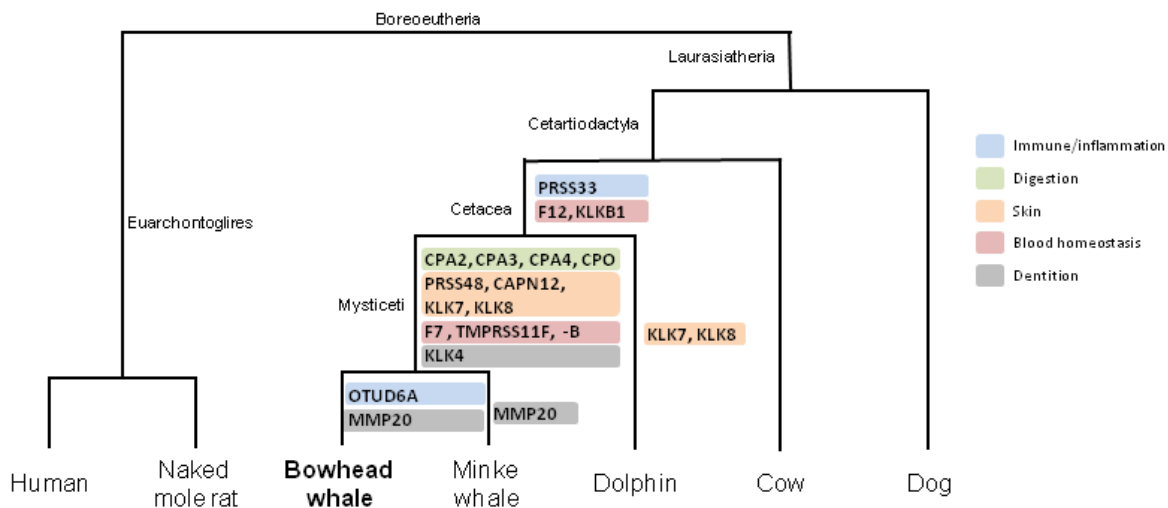
**Figure S4: Putative LAMTOR1 gene duplication**

ENSBTAG00000003001	MGCCYSENEDSDQDREERKLLDPSSPPTKALNGAEPNYHSLPSARTDEQALLSSILAKTASNIIDVSAADSQG
ENSOANG00000001786	---CKLTLPHPRQEREERKLLDPSSPPTKALNGTEPNYHSLPSARTDEQALLSSILAKTASNIIDVSAADSQG
ENSTTRG000000010763	MGCCYSENEDSDQDREERKLLDPSSPPTKALNGAEPNYHSLPSARTDEQALLSSILAKTASNIIDVSAADSQG
bmy_03663	MGCCYSENEDSDQDREERKLLDPSSPPTKALNGAEPNYHSLPSARTDEQALLSSILAKTASNIIDVSAADSQG
ENSMUSG000000030842	MGCCYSENEDSDQDREERKLLDPSSPTKALNGAEPNYHSLPSARTDEQALLSSILAKTASNIIDVSAADSQG
ENSCAFG00000005788	MGCCYSENEDSDQDREERKLLDPSSPPTKALNGAEPNYHSLPPTRTDEQALLSSILAKTASNIIDVSAADSQG
ENSG000000149357	MGCCYSENEDSDQDREERKLLDPSSPPTKALNGAEPNYHSLPSARTDEQALLSSILAKTASNIIDVSAADSQG
BACU019752G	MGRCYGSGNGDWDQDREERKLLDP--PPPKALNGAEPNYHSLPSARTDEQALLSSVLAKTAGNIIDVCASDSQG
bmy_21325	MGCCYSENEDSDQDREERKLLDPSSPPTKALNGAEPNYHSLPSASTDEQALLSSILAKTASNIIDVSAADSQG
ENSBTAG00000003001	MEQHEYMDRARQYSTRLAVLSSSLTHWKKLPPLPSLTSQPHQVLASEPVPFADLQ-----
ENSOANG00000001786	MEPHEYMDRARQYSTRLAVLSSSLTHWKKLPPLPSLTNQPHQVLASDPVPFADLQ-----
ENSTTRG000000010763	MEQHEYMDRARQYSTRLAVLSSSLTHWKKLPPLPSLTSQPHQVLASEPVPFADLQ-----
bmy_03663	MEQHEYMDRARQYSTRLAVLSSSLTHWKKLPPLPSLTSQPHQVLASEPVPFADLQ-----
ENSMUSG000000030842	MEQHEYMDRARQYSTRLAVLSSSLTHWKKLPPLPSLTSQPHQVLASEPIPFSDLQ-----
ENSCAFG00000005788	MEQHEYMDRARQYSTRLAVLSSSLTHWKKLPPLPSLTSQPHQVLASEPIPFSDLQ-----
ENSG000000149357	MEQHEYMDRARQYSTRLAVLSSSLTHWKKLPPLPSLTSQPHQVLASEPIPFSDLQQVRHPSAPAHPSHTAQGMA
BACU019752G	TEQHEGVDRARQCSTCLAVLSSSLTHWEKLPPLPSLSSQPHRVLASEPVPFADWQH-----
bmy_21325	TERHGYMDRARQYSTRLAVLSSSLTHWEKLPPLPSLTSQPHRVLASEPVLFADLQ-----
ENSBTAG00000003001	-----VSRIAAYAYSALSQIRVDAKEELVVQFGIP-----
ENSOANG00000001786	-----VSRIAAYAYSALSQIRVDAKEELVVQFGIP-----
ENSTTRG000000010763	-----VSRIAAYAYSALSQIRVDAKEELVVQFGIP-----
bmy_03663	-----VSRIAAYAYSALSQIRVDAKEELVVQFGIPX-----
ENSMUSG000000030842	-----VSRIAAYAYSALSQIRVDAKEELVVQFGIP-----
ENSCAFG00000005788	-----VSRIAAYAYSALSQIRVDAKEELVVQFGIP-----
ENSG000000149357	EGSPTLPQRRVSRIAAYAYSALSQIRVDAKEELVVQFGIPRHTGHTEKELVQLFQSTPCSQ
BACU019752G	-----VSRIAAYAGALSQIRVDAKEELVVQFGIPX-----
bmy_21325	-----VSRIAAYAGALSQIRVDAKEELVVQFGIPX-----

## Analysis of bowhead whale protease genes

Proteases form a diverse group of enzymes that share the ability to hydrolyze peptide bonds. The biological and pathological significance of this enzymatic activity has prompted the definition of the degradome as the complete repertoire of proteases in an organism<sup>1</sup>. From a genomic point of view, the degradome is highly attractive for several reasons. First, it is composed of a large number of genes. Thus, the human degradome includes about 600 protease genes, which represents almost 3% of the total annotated human protein-coding genes. Moreover, catalytic domains of proteases exhibit a high sequence diversity, which is further increased by the frequent attachment of auxiliary, non-proteolytic domains to the catalytic moieties<sup>2</sup>. Some of the protease genes have been shown to occur in genomic clusters, which is convenient for the study of short-term evolution. By contrast, most protease genes are randomly distributed throughout the annotated genomes. Therefore, the degradome forms a representative subset of the coding genome of a species. Notably, this structural diversity also reflects the multiple biological roles of proteases in every organism. Thus, beyond their obvious role in protein digestion, proteases also mediate regulatory processes through their ability to perform highly specific reactions of proteolytic processing, which have contributed to the acquisition of different functional capacities during evolution.

The comparison of the degradomes of the bowhead whale to those of minke whale, human and other mammals shows multiple events of gene loss in cetaceans and very few events of productive gene duplication. As expected, both whales share most of these genomic hallmarks, which probably reflect milestones in their evolution, including immune challenges, diet specialization, skin adaptation to the aquatic environment and changes in blood pressure and coagulation. Nevertheless, there are also some features specific for bowhead whale (Fig. S5).



**Figure S5. Genomic losses in the bowhead whale degradome. Each gene is depicted on the right side of the branch where each loss is inferred. Putative roles of each protease are shown in different colors.**

### Immunity and inflammation

The immune system and inflammatory pathways must respond to a very different environment in aquatic mammals compared to their terrestrial counterparts. In addition, there is a large and growing body of research on the influence of the immune system in the ageing process<sup>3</sup>. As long-lived mammals, whales, and particularly the bowhead whale, provide adequate models to understand the physiological adaptations that allow individuals to survive past their reproductive age<sup>4</sup>. Consistent with this, we have found several high-impact variants in proteases related to these functions in cetaceans. Thus, the cysteine protease **CASP12** (Fig. S6A), a modulator of the activity of inflammatory caspases, has at least one conserved premature stop codon in bowhead and minke whales. Interestingly, while this protease is conserved and functional in almost all of the terrestrial mammals, most human populations display deleterious variants<sup>5</sup>, presumably with the same functional consequences as the premature stop codons in whales. Human individuals who display the uninterrupted version of CASP12, as well as animal models simulating this variant, are more sensitive to infection and sepsis<sup>6,7</sup>. Related to this loss, we have found that one of the splicing forms of the immunoproteasome subunit **PSMB8**, a threonine protease, was pseudogenized through a frameshift mutation causing two premature stop codons in a common ancestor to baleen whales (Fig. S6B). The immunoproteasome is a modified form of the proteasome induced by interferon gamma which is important in MHC class I peptide display. Thus, while in most mammals there are two major splicing forms of this gene, both of them expressed in multiple tissues (<http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/av.cgi?db=35g&c=Gene&l=PSMB8>), baleen whales only have one. In humans, a missense mutation of **PSMB8** which would affect both major splicing forms, leads to an autoinflammatory syndrome with lipodystrophy<sup>8</sup>. Notably, **THOP1**, another modulator of MHC

class I peptide display<sup>9</sup>, is one of the most important targets of selection in cetaceans, with specific variants which we have confirmed in bowhead whale (Fig. S6C). Similarly, a bowhead whale-specific change could be the loss of *OTUD6A*, also known as *DUB2A*, which has a putative role in the innate immune system<sup>10,11</sup>. However, these results need independent confirmation, since complete losses can be mimicked by assembly artifacts. The serine protease *PRSS33* has been lost in cetaceans through two conserved premature stop codons (Fig. S6D). Notably, all known losses of this macrophage-specific gene in mammals are independent. Chimpanzees lost *PRSS33* through an Alu-mediated recombination mechanism<sup>12,13</sup>, whereas the orthologs in orangutans and rhesus monkeys show different premature stop codons<sup>14</sup>. Therefore, this protease has been independently lost in multiple mammals, including cetaceans, probably reflecting the need for quick evolution of the immune system in different circumstances. Finally, the haptoglobin cluster of serine proteases (*HP* and *HPN*) has been previously singled out as a target for selection in cetaceans<sup>15</sup>. Bowhead *HP* is not in fact an ortholog of human *HP*, but of both human *HP* and *HPR* after a primate-specific duplication. After adding human *HPR* and several additional mammalian sequences to the alignment, we have confirmed most of the cetacean-specific residue changes, with the exception of N259D, which is also an aspartic acid in dogs (ENSCAFP00000029992) (Fig. S6E). This result supports the hypothesis that *HP*, encoding an antioxidant and proangiogenic protein, has undergone selective pressure in cetaceans, as has also been shown in primates. Taken together, these events show that, similar to other mammalian species, selective pressure in cetaceans has been significant on proteins involved in the immune system. It is noteworthy that some of the cetacean targets of selective pressure have also been selected in primates, in spite of their very different environment.

### Coagulation and blood pressure control

Multiple coagulation factors, most of them from the S01 family of serine proteases, have been lost in bowhead and minke whales. One of these proteases, F12, has also been inactivated in dolphins, and therefore its loss probably occurred at an early stage of adaptation to the aquatic medium (Fig. S6F). Thus, all three orthologs show a change in the catalytic site of the protease which would yield an inactive protease. In the case of the whales, early stop codons suggest that the protein is not produced. In humans, a deficiency in F12 causes alterations in the coagulation process<sup>16</sup>. This shows one example of how adaptation to a new environment is sometimes driven through changes that may be harmful in the original circumstances, in a process known as Dobzhansky anomaly. A related serine protease gene, *KLKB1*, has also been pseudogenized in a common ancestor to both whales, and is not found in dolphins. Both F12 and *KLKB1* participate in the kinin-kallikrein system, with known roles in inflammation, blood pressure control, coagulation and pain. In fact, a genome association analysis has found variants of these serine proteases related to increased levels of vasoactive peptides<sup>17</sup>. Another protease involved in this system, MME or neprilysin, has been singled out as one of the preferential targets of selection in cetaceans<sup>15</sup>, with specific changes that we have also found in bowhead whale (Fig. S6G). Similarly, ACE2 and LNPEP, involved in the related renin-angiotensin system, show multiple cetacean-specific sites with functional consequences, which we have confirmed in bowhead whales<sup>15</sup>. Finally, the related serine proteases F7, *TMPRSS11F* and *TMPRSS11B* are pseudogenes in bowhead and minke whales, but seem to be functional genes in dolphins. These changes suggest that the mammalian potential for clotting and blood pressure are excessive in an aquatic environment, and these systems had to be modulated through the loss of proteases implicated in related proteolytic cascades.

## Digestive system

Several paralogues of carboxypeptidase A from the M14 family of metalloproteases have been pseudogenized in bowhead and minke whales. Thus, **CPA2** and **CPA3** show premature stop codons in bowhead whale (Fig. S6H). Most of these stop codons are conserved in the genome of the minke whale. However, the overall sequence of the predicted proteins is well conserved, which suggests that these pseudogenization events took place recently in a common ancestor. Consistent with this, dolphins show normal orthologs for each of the human CPA genes. The pattern of specific inactivation by point mutations instead of by gene loss might be related to the fact that all CPA-like genes are clustered in the genome. This mechanism might be related to the need to preserve CPA1 and CPA5 active. Both CPA1 and CPA2 are expressed mainly in pancreas and play an important role in protein digestion and absorption<sup>18</sup>. Therefore, the loss of CPA2 is likely to be related to the specialized diet of cetaceans. Supporting this hypothesis, we have also found conserved premature stop codons in the cetacean orthologs of CPO, an additional carboxypeptidase from the same family which is expressed in intestinal epithelial cells<sup>19</sup> (Fig. S6I). The specific evolution of proteases involved in the digestion of dietary proteins in cetaceans is further supported by the finding of five cetacean-specific sites in **ANPEP**, not present in other mammals<sup>15</sup>. **ANPEP** encodes a metalloprotease implicated in the final digestion of peptides generated from hydrolysis of proteins by gastric and pancreatic proteases<sup>20</sup>. The loss of CPA3 might be related to the same adaptive mechanism, since this enzyme is also found in pancreatic secretions<sup>21</sup>. Interestingly, CPA3 has also been studied in connection to the modulation of innate immune response and blood pressure<sup>22</sup>, which suggests that the loss of this protein might be involved in adaptation to the aquatic environment.

## Skin

Multiple kallikreins from the S01 family of serine proteases have been likewise pseudogenized in both bowhead and minke whales (Fig. S6J). Interestingly, two of the lost kallikreins, **KLK7** and **KLK8**, have been implicated in skin homeostasis<sup>23</sup> and are also absent in dolphins. While bowhead and minke whales show conserved premature stop codons in the predicted sequence of these genes, dolphins display premature stop codons at different positions, suggesting a case of converging molecular evolution. The specific loss of two genes through independent mechanisms strongly suggests that this is an important evolutionary event, which could be related to the adaptation of the mammalian skin to aquatic environments. In fact, KLK8 has been directly related to terminal differentiation and desquamation of the stratum corneum, the outmost layer of the skin in mammals<sup>24</sup>. An additional skin-specific but not so well characterized serine protease, **PRSS48**, has been similarly lost in both whales. Finally, **CAPN12**, a cysteine protease preferentially expressed at the cortex of the hair follicle<sup>25</sup>, has been lost in bowhead and minke whales (Fig. S6K). According to these observations, some of the differential characteristics of cetacean skin, like their parakeratotic stratum corneum with incomplete keratinization or its renewal through flaking rather than desquamation, might be related to the loss of several proteases<sup>26,27</sup>. Also noteworthy is the duplication of the cysteine protease **UCHL3** through a retrotranscription-mediated process. While this duplication seems to have happened in a common ancestor to mysticetes, only the genome of the bowhead whale shows a complete, putatively functional coding sequence for a **UCHL3**-like protease. This protease has been linked to adipogenesis, which suggests that this duplication might be related to the adaptation to the harsh arctic climate where this whale thrives.



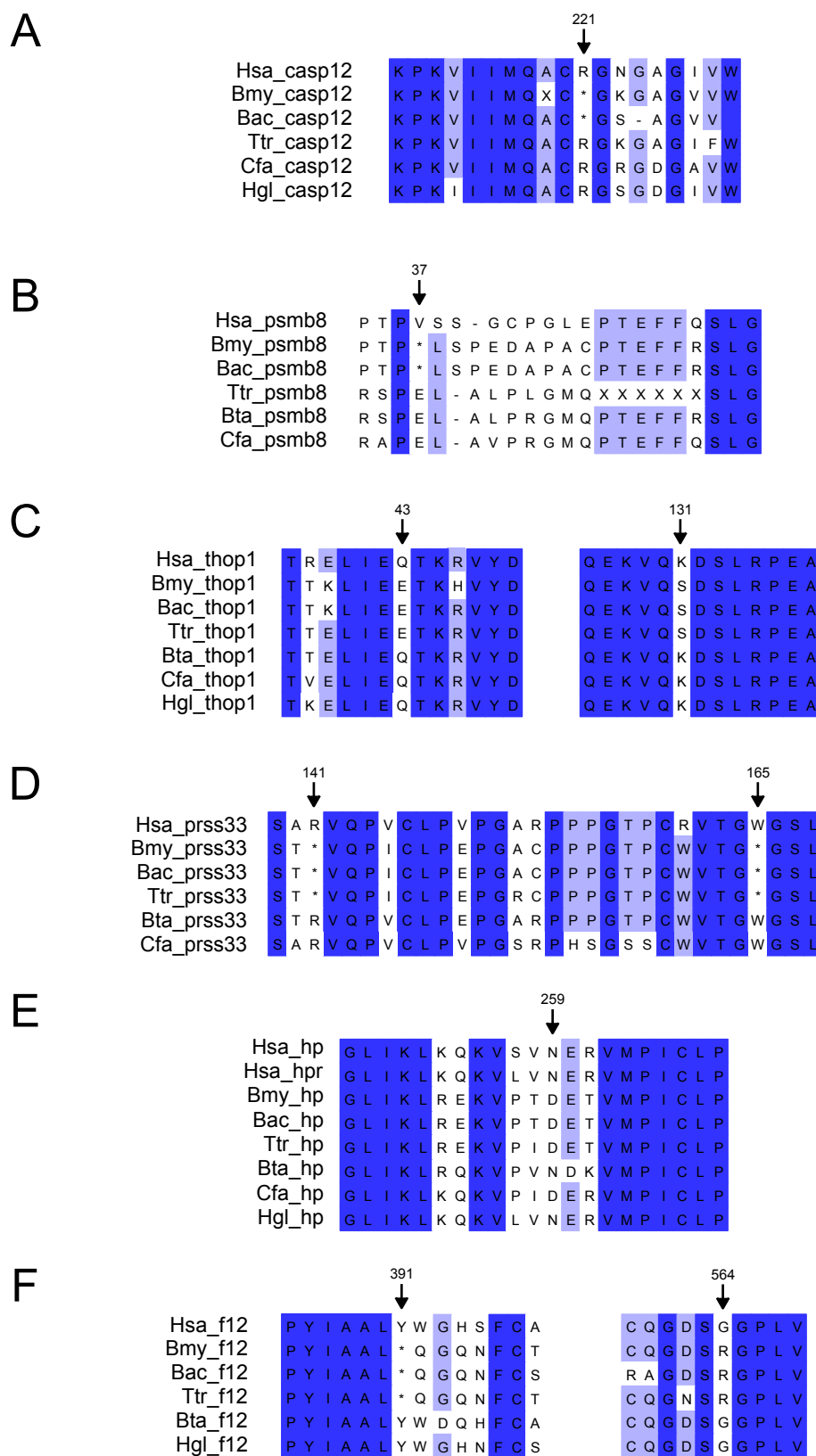
## Dentition

**KLK4** was pseudogenized through a frameshift mutation in a common ancestor to both whales, but not in dolphins (Fig. S6L). This protease is involved in dental enamel formation, and its pseudogenization in mammals, in concert with that of the metalloprotease **MMP20**, leads to amelogenesis imperfecta in mammals<sup>28,29</sup>. The loss of **MMP20** in mysticetes has been previously documented<sup>15,30</sup>. We have found that the pseudogenization of bowhead whale **MMP20** has followed a different path to that of minke whale (Fig. S6M). Thus, unlike the minke whale ortholog, the predicted open reading frame of bowhead whale **MMP20** contains no early stop codons. Instead, the initiation methionine has been mutated to an isoleucine, which is expected to hamper translation of an active protein. Even if a different methionine residue were used as initiator, the resulting protein would lose its signal peptide, which is necessary for its extracellular function. Therefore, the loss of both **KLK4** and **MMP20** is likely to be related to the loss of teeth in the suborder Mysticeti. Even though an insertion of a SINE element has been proposed as a common mechanism for the loss of **MMP20** in mysticetes, our data support different independent mechanisms in several of the species.

## References

- 1 Quesada, V., Ordonez, G. R., Sanchez, L. M., Puente, X. S. & Lopez-Otin, C. The Degradome database: mammalian proteases and diseases of proteolysis. *Nucleic Acids Res* **37**, D239-243 (2009).
- 2 Lopez-Otin, C. & Overall, C. M. Protease degradomics: a new challenge for proteomics. *Nat Rev Mol Cell Biol* **3**, 509-519 (2002).
- 3 Lopez-Otin, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The hallmarks of aging. *Cell* **153**, 1194-1217 (2013).
- 4 Sierra, E. *et al.* Muscular senescence in cetaceans: adaptation towards a slow muscle fibre phenotype. *Sci Rep* **3**, 1795 (2013).
- 5 Fischer, H., Koenig, U., Eckhart, L. & Tschachler, E. Human caspase 12 has acquired deleterious mutations. *Biochem Biophys Res Commun* **293**, 722-726 (2002).
- 6 Saleh, M. *et al.* Differential modulation of endotoxin responsiveness by human caspase-12 polymorphisms. *Nature* **429**, 75-79 (2004).
- 7 Yeretssian, G. *et al.* Gender differences in expression of the human caspase-12 long variant determines susceptibility to *Listeria monocytogenes* infection. *Proc Natl Acad Sci U S A* **106**, 9016-9020 (2009).
- 8 Kitamura, A. *et al.* A mutation in the immunoproteasome subunit PSMB8 causes autoinflammation and lipodystrophy in humans. *J Clin Invest* **121**, 4150-4160 (2011).
- 9 York, I. A. *et al.* The cytosolic endopeptidase, thimet oligopeptidase, destroys antigenic peptides and limits the extent of MHC class I antigen presentation. *Immunity* **18**, 429-440 (2003).
- 10 Kayagaki, N. *et al.* DUBA: a deubiquitinase that regulates type I interferon production. *Science* **318**, 1628-1632 (2007).
- 11 Meenhuis, A., Verwijmeren, C., Roovers, O. & Touw, I. P. The deubiquitinating enzyme DUB2A enhances CSF3 signalling by attenuating lysosomal routing of the CSF3 receptor. *Biochem J* **434**, 343-351 (2011).
- 12 Puente, X. S., Gutierrez-Fernandez, A., Ordonez, G. R., Hillier, L. W. & Lopez-Otin, C. Comparative genomic analysis of human and chimpanzee proteases. *Genomics* **86**, 638-647 (2005).
- 13 Johnson, M. E. *et al.* Recurrent duplication-driven transposition of DNA during hominoid evolution. *Proc Natl Acad Sci U S A* **103**, 17626-17631 (2006).

- 14 Locke, D. P. *et al.* Comparative and demographic analysis of orang-utan genomes. *Nature* **469**, 529-533 (2011).
- 15 Yim, H. S. *et al.* Minke whale genome and aquatic adaptation in cetaceans. *Nat Genet* **46**, 88-92 (2014).
- 16 Renne, T., Schmaier, A. H., Nickel, K. F., Blomback, M. & Maas, C. In vivo roles of factor XII. *Blood* **120**, 4296-4303 (2012).
- 17 Verweij, N. *et al.* Genome-wide association study on plasma levels of midregional-proadrenomedullin and C-terminal-pro-endothelin-1. *Hypertension* **61**, 602-608 (2013).
- 18 Szmola, R. *et al.* Chymotrypsin C is a co-activator of human pancreatic procarboxypeptidases A1 and A2. *J Biol Chem* **286**, 1819-1827 (2011).
- 19 Lyons, P. J. & Fricker, L. D. Carboxypeptidase O is a glycosylphosphatidylinositol-anchored intestinal peptidase with acidic amino acid specificity. *J Biol Chem* **286**, 39023-39032 (2011).
- 20 Fairweather, S. J., Broer, A., O'Mara, M. L. & Broer, S. Intestinal peptidases form functional complexes with the neutral amino acid transporter B(0)AT1. *Biochem J* **446**, 135-148 (2012).
- 21 Whitcomb, D. C. & Lowe, M. E. Human pancreatic digestive enzymes. *Dig Dis Sci* **52**, 1-17 (2007).
- 22 Pejler, G., Knight, S. D., Henningsson, F. & Wernersson, S. Novel insights into the biological function of mast cell carboxypeptidase A. *Trends Immunol* **30**, 401-408 (2009).
- 23 Kishibe, M. *et al.* Kallikrein 8 is involved in skin desquamation in cooperation with other kallikreins. *J Biol Chem* **282**, 5834-5841 (2007).
- 24 Kuwae, K. *et al.* Epidermal expression of serine protease, neuropsin (KLK8) in normal and pathological skin samples. *Mol Pathol* **55**, 235-241 (2002).
- 25 Dear, T. N., Meier, N. T., Hunn, M. & Boehm, T. Gene structure, chromosomal localization, and expression pattern of Capn12, a new member of the calpain large subunit gene family. *Genomics* **68**, 152-160 (2000).
- 26 Spearman, R. I. The epidermal stratum corneum of the whale. *J Anat* **113**, 373-381 (1972).
- 27 Haldiman, J. T. *et al.* Epidermal and papillary dermal characteristics of the bowhead whale (*Balaena mysticetus*). *Anat Rec* **211**, 391-402 (1985).
- 28 Wang, S. K. *et al.* Novel KLK4 and MMP20 mutations discovered by whole-exome sequencing. *J Dent Res* **92**, 266-271 (2013).
- 29 Yamakoshi, Y. *et al.* Enamel proteins and proteases in Mmp20 and Klk4 null and double-null mice. *Eur J Oral Sci* **119 Suppl 1**, 206-216 (2011).
- 30 Meredith, R. W., Gatesy, J., Cheng, J. & Springer, M. S. Pseudogenization of the tooth gene enamelysin (MMP20) in the common ancestor of extant baleen whales. *Proc Biol Sci* **278**, 993-1002 (2011).



**Figure S6. Alignments for selected sequences in cetaceans and other mammals.** Numbers correspond to human proteases. *Hsa*, human; *Bmy*, bowhead whale; *Bac*, minke whale; *Ttr*, bottlenose dolphin; *Bta*, cow; *Cfa*, dog; *Hgl*, naked mole rat (continued).

[illegible]

Species	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Hsa_cpo	S	E	K	Y	K	E	V	V	T	Q	H	F	L	G	V	T
Bmy_cpo	R	E	K	Y	T	E	V	V	T	*	H	F	L	G	M	T
Bac_cpo	R	E	K	Y	T	E	V	V	T	*	H	F	L	G	M	T
Ttr_cpo	R	E	K	Y	T	E	V	V	T	*	H	F	L	G	M	T
Bta_cpo	R	E	K	Y	T	E	V	V	T	Q	H	F	L	G	M	T
Cfa_cpo	S	E	K	Y	A	G	V	A	T	Q	H	F	L	G	M	T

221

Hsa_capn12	I	I	M	Q	A	C	R	G	N	G
Bmy_capn12	I	I	M	Q	X	C	*	G	K	G
Bac_capn12	I	I	M	Q	A	C	*	G	K	G
Ttr_capn12	I	I	M	Q	A	C	R	G	K	G
Cfa_capn12	I	I	M	Q	A	C	R	G	R	G
Hgl_capn12	I	I	M	Q	A	C	R	G	S	G

Hsa\_klk4 V S G W G L L A N G R M  
Bmy\_klk4 P A G V G - \* R M G R L  
Bac\_klk4 P A G V G - \* R M G R L  
Ttr\_klk4 V S G W G R L K N G R L  
Cfa\_klk4 V S G W G Q L I D G R Q  
Hgl\_klk4 V S G W G R L A N G G L

156  
↓

1

114

Hsa\_mmp20  
Bmy\_mmp20  
Bac\_mmp20  
Ttr\_mmp20  
Bta\_mmp20  
Cfa\_mmp20  
Hgl\_mmp20

Y R L F P G E P K W K K  
Y C L F A G E S K W K K  
Y R L F P G \* P K W K K  
Y R L F P G E P K W K K  
Y R L F P G E P K W K K  
Y R L F P G E P K W K K  
Y R L F P G E P K W E K

108